A Cognitive Approach to Multimodal Attention

Raúl Arrabales, Agapito Ledezma and Araceli Sanchis

***Abstract*—An adaptive attention mechanism is a requirement when an autonomous robot has to deal with real world environments. In this paper we present a novel cognitive architecture which enables integrated and efficient filtering of multiple modality sensory information. The proposed attention mechanism is based on contexts that determine what sensorimotor data is relevant to the current situation. These contexts are used as a mean to adaptively select constrained cognitive focus within the vast multimodal sensory space. In this framework, the focus of attention can be directed to meaningful complex percepts, thus allowing the implementation of higher cognitive capabilities. Sonar, contact, and visual sensory modalities have been used in the perception process, and the motor capability of the physical agent is provided by a differential wheel drive system. The testing of this artificial attention approach, carried out initially in the domain of counterpart recognition and chasing, has demonstrated both a great decrease in computation power requirements and ease of multimodal integration for cognitive representations.**

***Index Terms*—Physical agents, Attention, cognitive modeling, mobile robotics.**

I. Introduction

DESIGNING an autonomous control system for a mobile robot implies a decision on what inputs will be handled and what repertory of actions can be executed at any given time. The option of considering all the available sensory information as input for the core control of the robot is usually both unnecessary and extremely expensive in computational terms. Analogously, not all possible robot behaviors are appropriate at any given time. Instead of considering all physically plausible behaviors, the robot control system should take into account its current situation and assigned mission in order to build a shorter list of eligible behaviors. A simplistic definition of artificial attention can be drawn from the problem described above. Hence, let us say that an efficient artificial mechanism for attention would solve the problem of filtering relevant sensory information and selecting relevant behaviors.

According to the former definition, we need to specify what relevant means in terms of implementing an efficient attention mechanism. Relevant sensor data and relevant behaviors are those that could be both useful to accomplish the mission and adapted to the world in which the robot is situated. Attention has been typically applied to artificial vision systems taking the human visual attention mechanisms and its related eye movement control (foveation) as inspiration [1]. Visual attention has been extensively applied in robotics, e.g. [2]. However, much less effort has been put in pure multimodal attention mechanisms [3]. Usually attention mechanisms for robots focus in great degree on visual sensory information; nevertheless, some salient examples incorporate data from other sensors in the attention mechanism. For instance, laser range finders [4]. In this work we present a purely multimodal attention mechanism, which permits a straightforward and graceful inclusion of new additional sensors of different modalities. The proposed mechanism for multimodal integration is not only intended to exclusively serve agent's attention capability, but also to provide a rich, complex, and coherent percept representation that can be directly used by other cognitive functions like associative learning and decision making.

Fig. 1. Mobilerobots Pioneer 3 DX robot.

Currently, sonar, contact, and vision modalities have been already incorporated in the perception subsystem. The actuators subsystem consists exclusively on a two-motor set forming a single differential wheel drive that provides the required indoor mobility. The testing platform is based on a Mobilerobots Pioneer 3 DX robot (P3DX) equipped with an onboard laptop computer, frontal centered fixed single camera, eight-transducer frontal sonar ring, and frontal and rear bumper rings (see Fig. 1). A counterpart recognition and chasing task has been selected as preliminary testing domain for the proposed cognitive attention mechanism. Both simulated and real environments have been setup as described below. In the simplest scenario, two P3DX robots are used: *P3DX-Chaser* is the robot running the autonomous control architecture which implements the proposed attention mechanism, and *P3DX-Target* is a similar robot base tethered or remotely controlled by a human. The mission consigned to *P3DX-Chaser* is to keep heading towards *P3DX-Target* maintaining a safe constant distance between the two robots. In order to accomplish the chasing goal, *P3DX-Chaser* has to pay attention to complex percepts such as *"a moving target which is a Pioneer 3 DX robot"*, while ignoring other percepts which are irrelevant to current mission. Being able to deal with such complex percepts when focusing attention is one of the main goals of this work in cognitive artificial attention.

Raúl Arrabales is with University Carlos III of Madrid.
E-mail: rarrabal@inf.uc3m.es
Agapito Ledezma is with University Carlos III of Madrid.
E-mail: ledezma@inf.uc3m.es
Araceli Sanchis is with University Carlos III of Madrid.
E-mail: masm@inf.uc3m.es

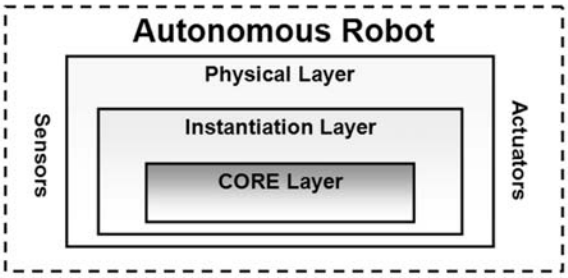

Fig. 2. CERA Control Architecture Layers.

In the next sections we discuss the implementation of an attention mechanism able to fulfill the requirement of selecting relevant sensorimotor information. Section II provides an introduction to the software architecture and how the attention mechanism is incorporated into a layered control system. Section III covers the definition of the attentional contexts that are used to form sets of sensory and motor data. Section IV is dedicated to explain how the proposed mechanism allows the integration of different modality sensory information into the same context. Section V illustrates the application of the proposed technique to the domain of counterpart recognition and chasing. Finally, we conclude in section VI with a discussion of the benefits and possible areas of application of the attention mechanism in the field of cognitive robotics.

II. Architecture Overview

Typically, autonomous robot control architectures are structured in layers. Each layer usually represents a different level of control, from lower reactive levels to higher deliberative levels. The proposed attention mechanism has been integrated into a three level control architecture called CERA (Conscious and Emotional Reasoning Architecture). CERA is composed of a lower level, called Physical Layer, a mission specific level, called Instantiation Layer, and a higher level, called Core Layer, where higher cognitive functions are implemented (see Fig. 2). The details about CERA are discussed elsewhere [5].

A number of processing steps that take place within this architecture can be identified as specifically belonging to the attention mechanism. Concretely, if we look at the perception cycle, the following steps are performed (see Fig. 3):

- Sensory data is acquired by sensors (for instance, a bump panel contact is reported).
- Contextualization parameters (like relative position vectors and timestamps) are calculated for each perceived object or event.
- Sensor Preprocessors build single percepts using both sensory data and their associated contextualization parameters.
- Groups of single percepts showing contextual affinity are eventually combined into complex multimodal percepts.

CERA Physical Layer provides the required functionality in order to interface with the robot hardware. In other words, it provides access to sensors and actuators. Additionally, as the CERA architecture has been designed to host the proposed attention mechanism, the physical layer is also in charge of

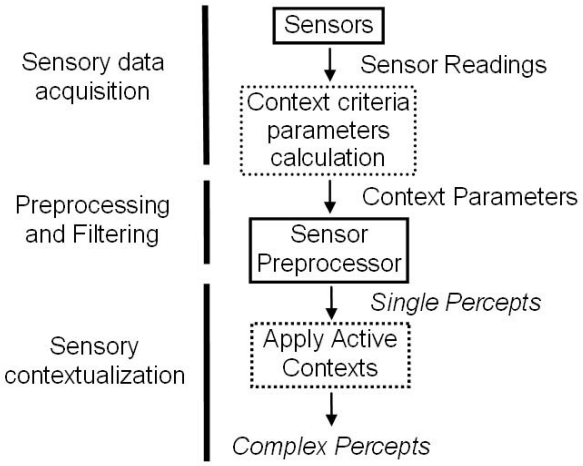

Fig. 3. Perception cycle overview.

calculating the contextual parameters of percepts and actions. From the point of view of the attention mechanism, the CERA Physical Layer is the domain of single percepts and simple actions. As the Physical Layer is specific to a given hardware it has to be changed or adapted if the underlying physical robot is replaced by a significantly different model. The composition of percepts and actions forming complex percepts and complex actions takes place in the CERA Instantiation Layer. This is the place where mission-specific contexts are to be applied, and therefore mission-specific complex percepts and behaviors are generated. As the Instantiation Layer is designed specifically for a given problem domain it can be replaced by a different problem instantiation without changing the existing Physical and Core layers. Finally, the CERA Core Layer is where a machine consciousness model is implemented based on several modules that represent higher cognitive functions. One of these functions related to consciousness is attention.

The attention module implemented in the Core Layer is designed to activate the most appropriate contexts at any given time, i.e. an attentional bias is induced from the Core Layer enforcing particular contexts. Complex percepts that are obtained in the perception cycle depend on the active contexts established by the Core Layer. Therefore, at any given time, the robot can only perceive those objects or events that are relevant to the mission (top-down attentional bias). However, a mechanism for bottom-up attention is always in place, so critical single percepts like bumper contact notifications are not ignored. One of the benefits of integrating the attention mechanism into a layered control system, where priorities for perceptions and actions can be established, is that the implementation of a combination of top-down and bottom-up attentional bias can be naturally enabled.

III. Definition of Attentional Contexts

Our proposed artificial attention mechanism is inspired in the concept of context as defined in the Global Workspace Theory (GWT) [6]. The GWT is a cognitive account for consciousness, and therefore it covers attention as a key characteristic of conscious beings. However, the GWT do not

provide any algorithmic description of attention but just a metaphorical explanation. A theater spotlight simile is used to represent the focus of consciousness. This spotlight illuminates only a small part of the scene, which is considered the conscious content of the mind. The scene is actually built upon the subject's working memory. The movement of the spotlight, i.e. the selection of contents that will be used for volition and action, is directed by unconscious contextual systems. The aim of the work described in this paper is to design and test an implementation of such contextual systems, which are able to adaptively direct attention toward the interesting areas of the robot sensorimotor space.

From the point of view of perception, contexts are sets of percepts retrieved from the sensors. Percepts are considered the minimal information units obtained by the robot sensory machinery [5]. Therefore, a sensory context can be used to build a complex percept composed of related single percepts. From the point of view of behavior, contexts define sets of actions available for execution. Hence, we can define behavioral contexts as possible compositions of related actions. In order to generate an efficient robot behavior, both sensory contexts and behavioral context have to be adaptively generated.

A. Visual Field Segmentation

The first stages in visual sensor data processing are concerned with attentional context definition. Concretely, instead of applying a full preprocessing task to the entire image captured by the camera sensor, each incoming frame is fragmented into smaller regions. Subsequently, only one selected fragment (foveal region) is further processed, thus reducing to a great extend the processor requirements of visual sensor preprocessor. Additionally, as explained below, this strategy allows the robot to focus attention in specific visual regions also in further processing stages. Nevertheless, before the preprocessing stage, when context criteria are evaluated, all visual data packages (frame segments) are equally processed.

It is known that this strategy is similar to the way human visual system processes the foveal region, which is much richer in resolution and detail than retinal periphery. Humans use the fovea to fixate on an object and specifically process its image while maintaining a much less demanding process for peripheral regions [7]. This very same strategy has also been successfully applied in other artificial systems, e.g. [8].

B. Context Criteria

We have designed the process of context formation as the application of predefined criteria in order to calculate the degree of relation between the potential elements of a given context. Basically, a context should be constructed in a way that it can become a meaningful representation of the reality, i.e. the interplay between agent and situation must be enforced by a proper definition of both sensory and behavioral contexts. The very basic factors that need to be considered in the correct representation of robot situation in the world are time and location. Nevertheless, other factors can be considered depending on the problem domain and internal state representation richness. In the work described here, color and movement properties have been considered as additional criteria; therefore, four criteria have been used for context formation in the experiments described below.

The time criterion refers to the exact moment at witch a stimulus is perceived. Therefore, it should be taken as an important criterion to relate one percept to another. Given that different sensors and their associated device drivers can take different time intervals to process the sensory information, a mechanism for time alignment is required. It has been demonstrated that such a time alignment mechanism is present in biological brains [9][10]. Although visual and auditory stimuli are processed at different speeds, the time gap between different processed signals, whose physical originators were acquired at the same time, is automatically removed by the brain [11]. An analogous artificial mechanism has been implemented in the proposed architecture.

Location is another fundamental criterion for context formation as the representation of the position of objects in the world is a requirement for situatedness. Furthermore, the location of an object relative to the robot body (or any other reference frame) is required for generating adaptive behaviors. The relative location of any element in the sensory world is necessary for the integration of complex percepts; additionally, it allows the selection of a given direction of attention toward the most relevant places. The presence of space coding neurons and the use of reference frames (like somatotopic or head-centered) has been demonstrated in the mammal brain [12][13].

In a world where color patterns can be associated with particular objects, this property of entities should be taken into account. Similarly, some objects are mobile while others remain static; consequently, movement is a property that should also be considered as criterion for relevant context formation. Particularly, the task of counterpart recognition has been simplified in the research under discussion by characterizing other peer robots as autonomously moving red and black objects. The presence of specialized areas for color and movement detection has been demonstrated in human's brain visual cortex [14].

Following the principles presented above, we have used time, location, color, and motion as fundamental contextualization criteria for the formation of:

- Sensory contexts as composition of single percepts (complex percepts), and
- behavioral contexts as composition of simple actions.

In order to generate these contexts, both single percepts (which are built from data packages obtained from sensors) and simple actions (which are defined as part of the robot control system) are required to incorporate estimated time, location, motion, and color parameters (see Fig. 4). Motion properties could be obviously derived from time and location parameters; however, we have decided to use a natively visual motion detection approach in which motion properties are directly obtained from visual input analysis. In our proposed architecture there are specialized modules designed to calculate time, color, motion, and location parameters: the Timer module maintains a precision clock (less than 1 millisecond resolution) that represents the robot's age, the Proprioception

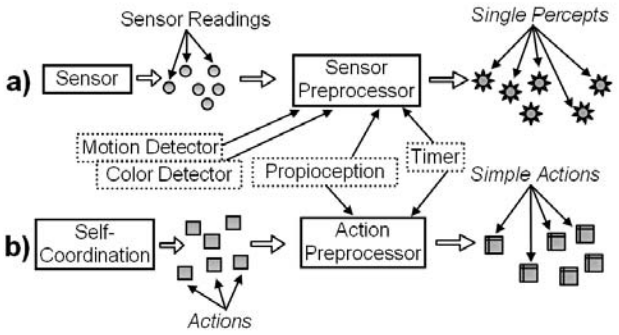

Fig. 4. Creation of single percepts and simple actions

module maintains all the required information to calculate the exteroceptive sensors position. This information is necessary to estimate the relative location of an object or event detected by an exteroceptive sensor. A Color Detection module is in charge of providing a color histogram representation associated to visual data retrieved by the camera. Similarly, a Motion Detection module continuously calculates the differences between the last data retrieved by the camera and current visual data.

Time, location, color, and motion parameters provided by the Timer, Proprioception, and Color and Motion Detector modules are used by the preprocessor modules in charge of generating single percepts and simple actions. A Sensor Preprocessor takes a given sensor reading as input, then calculates the relative position of the source of the reading and the instant when it took place using the information provided by the Timer and Proprioception. In case of visual sensor readings, also color and motion detectors are activated and histogram and motion vectors are calculated. Finally, the sensor preprocessor creates a single percept packing together the proper sensor reading with its contextualization information. The Action Preprocessor takes as input an action generated by the Self-Coordination module (this module and the way it works is described elsewhere [5]), and applies the same approach as in the Sensor Preprocessor in order to build the Simple Action representations.

More parameters should be added to single percepts if other contextualization criteria are to be applied. In the work described in the present paper, the following parameters have been used:

- **Timestamps**: two different timestamps are recorded in single percepts. The first timestamp is set when the sensory data is collected from the sensor. Usually this timestamp is directly assigned by the sensor hardware and retrieved in the control system thought the sensor driver. The second timestamp is set when the percept is actually used in the control system. The time span between these two timestamps can be significant when a sensor is incessantly notifying readings and there is not enough onboard processing power to dispatch all the incoming data. Actually, the time span value can be used to discard too old sensory data which is not significant to the current robot state. Similarly, two timestamps are logged in the case of simple action. The first one is set when the simple action is created and enqueued in the control system. The second timestamp is set when the action enters the core execution cycle, i.e. when the action is actually dequeued and dispatched (begins physical execution). The time span between these two timestamps can be used to detect delays in the execution queue and eventually abort too old actions.
- ***J-Index***: for the representation of the location parameter of both single percepts and simple actions we have decided to use the robot body center of mass as reference frame. The term J-Index refers to a structure able to represent or map the relative position of an object or event within a biological brain [15]. We have adapted and enhanced the original definition of the J-Index representation with the aim of representing both the relative position and relative dimensions of the object. Hence, our J-Indexes are implemented as a composition of several n-dimensional vectors. The main vector is called the j referent vector, and is used to calculate the relative position of the geometrical center of the percept's source or the geometrical target of an action. Depending on the nature of the sensor that is reporting the sensory data, more positional vectors can be calculated in order to estimate the size of the percept (examples for sonar range finder, camera, and bump panel arrays are described below).
- ***Color Histogram***: Each data package provided by the visual sensor (corresponding to a frame segment) is assigned a color histogram, where the frequency of image color components is represented. Obviously this parameter can only be set for visual sensory information. Any other more demanding visual processing concerned with color, like texture recognition is not defined as contextual parameter because it will be limited to the scope of foveal region (when the robot is fixating on a particular object); therefore, it must be part of the sensor preprocessing task.
- ***M-Index***: The result of the application of the motion detection module on an incoming visual data package is a movement vector called M-Index, whose value is zero when no movement has been detected. Although motion could be detected using other sensory modalities, like sonar, we have decided to use only vision for the time being. Nevertheless, when complex percepts are built, the robot own movement is taken into account to calculate the relative motion of observed objects.

The timestamp parameters are easily acquired using the robot's control system precision timer. However, the J-Index parameters require more elaboration, particularly in the case of movable sensors. In the case discussed here, we have used P3DX robots (see Fig. 5a) with fixed position sensors: a frontal sonar array (see Fig. 5c) and frontal and rear bump panels (see Fig. 5b). In the experiments that we have carried out so far, J-Indexes have been calculated for sonar readings, bump panels contact and release notifications, and visual segments. The J-Indexes are calculated as a function of the transducer (fixed) position and orientation (relative to the robot front).

Although the J-Index parameter can be primarily repre-

sented by a three-dimensional vector, for the task of following a counterpart robot in a flat surface, a two-dimensional j referent vector can be considered, where (X,Z) = (0,0) represents the subjective reference frame of the robot (see Fig. 5b and 5c). Nevertheless, a Y coordinate (height) is usually calculated even though it is not used.

The calculation of the j referent vector is different depending on the sensor. In the case of bump panels, as they are located at angles around the robot (see Fig. 5b), the j referent vector is calculated using (1). Where, BR is the bump panel radius, i.e. the distance from the center of mass of the robot to the bumper contact surface (see Fig. 5b). BA is the bump panel angle to the front of the robot (Pioneer 3 DX bump panels are located at angles $-52°$, $-19°$, $0°$, $19°$, and $52°$). BH is the height at which the bumpers are mounted.

$$j = (X, Y, Z) = \begin{pmatrix} BR * Cos(BA) \\ BH \\ BR * Sin(BA) \end{pmatrix} \quad (1)$$

Additionally, two more vectors are calculated to be associated to a bumper percept: the left-j referent and the right-j referent (see Fig. 6). These two vectors represent the dimensions of the percept (the width assigned to the collision).

In order to calculate the j referent vector corresponding to a given sonar reading, (2) is used. Note that the calculation of j referent vectors is dependent on the type of sensor being considered.

$$j = (X, Y, Z) = \begin{pmatrix} (R + SR) * Cos(SA) \\ SH \\ (R + SR) * Sin(SA) \end{pmatrix} \quad (2)$$

Where, R is the maximum range measured by the sonar transducer, SR is the distance from the center of mass of the robot to the sonar transducer, and SA is the angle at which the particular sonar transducer is located. Note that sonar transducers are located at angles $-90°$, $-50°$, $-30°$, $-10°$, $10°$, $30°$, $50°$, and $90°$ to the front of the robot (see Fig. 5c). Therefore, each transducer is able to measure the free space available within a three-dimensional $15°$ wide cone (this cone aperture corresponds to the SensComp 600 transducer).

Taking into account that the ultrasonic beams emitted by the sonar transducers take the form of a symmetric three-dimensional cone, at least one additional j referent vector has to be calculated in order to estimate the dimensions of the single transducer sonar percept, i.e. the open space perceived in front of that particular sonar transducer. The main j referent vector calculated using (2) represents the cone bisector. Additionally, two more vectors: the left-j referent vector and right-j referent vector represent the lateral 2D boundaries of the percept (see Fig. 7). The representations of J-Indexes for both sonar and bumpers have been designed as described above with the aim of implementing an attention algorithm. Although some of the calculated reference vectors are expendable, they are useful to pre-calculate the regions of the world affected by a given percept. Besides, this representation is also particularly useful for the subsequent task of counterpart robot chasing.

In the case of visual sensory information, each segment is assigned a j referent vector which corresponds to the relative

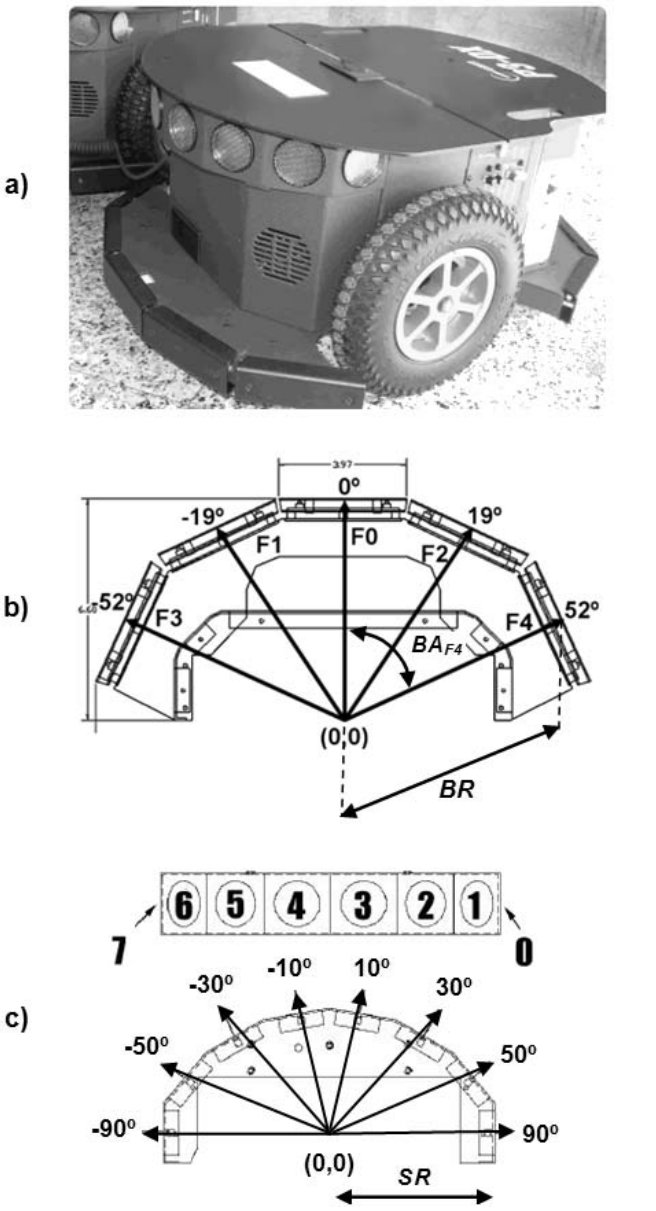

Fig. 5. MobileRobots Pioneer 3 DX Robot, frontal bumper panel, and frontal sonar ring.

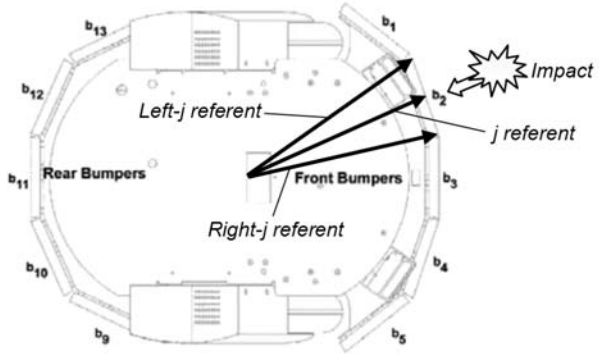

Fig. 6. Vectors calculated to build the J-Index of a single bump panel contact percept.

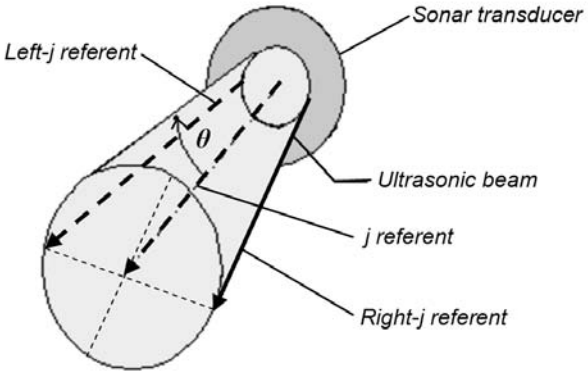

Fig. 7. Vectors calculated to build the J-Index of a single sonar transducer percept.

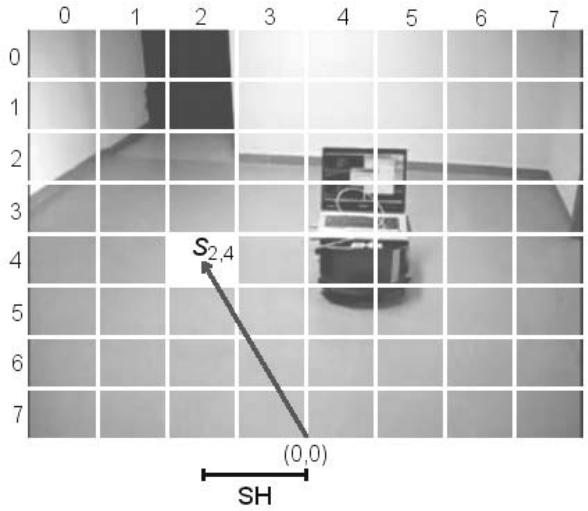

Fig. 8. J referent vector of a segment of visual sensory data.

location of the geometrical center of that particular segment within the visual field. As the orientation of the camera is fixed, it is straightforward to estimate the relative X coordinate (left / right position relative to the robot) of the corresponding percept, being SH the distance from the optical vertical axis of the camera (center of the field of view) to the center of the segment. Fig. 8 depicts an example of segmented visual input in which the visual field has been divided into 64 smaller regions and the referent vector for segment S_{24} is calculated. Estimating the distance to the visual percept is a different matter. Usually a stereo vision system is used. Having just one camera, a pinhole model could be applied. Nonetheless, in this work distance to objects is provided exclusively by sonar percepts.

Fig. 9 shows an example where j referent vectors are calculated only for those segments in which any saliency has been detected. In this case, as the goal is to follow a red counterpart robot, two segments where the color histogram presents a salient frequency of red have been selected. At this point, when the sensor preprocessor is building single percepts from visual input, a foveal region is to be selected, and the rest of the image is discarded and not taking part

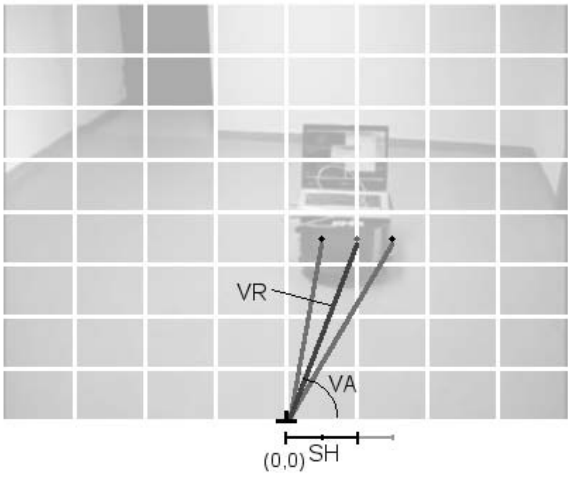

Fig. 9. J-Index of a single visual percept.

in any further processing. This means that no single percepts are built with visual information outside the simulated fovea. The foveal region has to be kept small, therefore, a maximum of four contiguous segments are considered to form a single visual percept. The J-Index of a single percept that has been formed as a combination of contiguous segments is calculated by adding a new j referent vector pointing to the geometrical center of the set of segments (see Fig. 9). Concretely, (3) is used to calculate the main j referent vector of the visual single percept, where CH is the relative height at which the camera is located, VR is the distance from the optical axis origin to the center of the percept, and VA is the angle relative to the optical horizontal axis ($SH = VR * Sin(VA)$).

$$j = (X, Y, Z) = \begin{pmatrix} VR * Sin(VA) \\ CH \\ ? \end{pmatrix} \quad (3)$$

Note that the calculation of all context criteria parameters is rather quick, and no complex processing is carried out at this level of the architecture. One of the advantages of having an attention mechanism is the processing power saving, and this principle is preserved by keeping simple context criteria parameters. When more processing is required in order to build complex percepts and apply inference rules, this is uniquely done using a reduced subset of the sensory space, which has been already selected by the application of a given context.

C. Actions Context Composition

As both Single Percepts and Simple Actions include the basic time and location contextualization parameters (timestamps and J-Indexes) it is straightforward to calculate similarity distances between them. Other specific sensory parameters, like color, are used exclusively with single percepts. Therefore, contexts can be generally defined based on the dimensions of relative time and relative location, and also specifically defined for some sensory modalities using other specific parameters. Each sensory context is used to build a representation structure called complex percept (see Fig. 10a). Complex percepts enclose the required information to represent the meaning of the

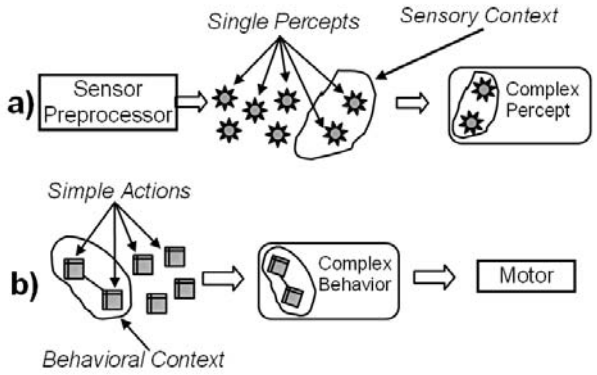

Fig. 10. Formation of complex percepts and complex behaviors.

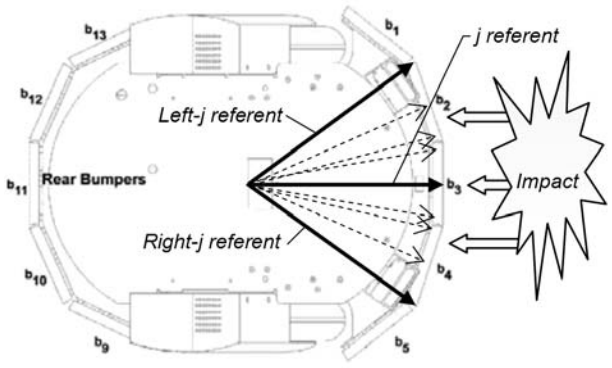

Fig. 11. Vectors calculated to build the J-Index of a complex bumper contact percept.

associated sensory context as required by the subsystems of the autonomous control system. As behavioral contexts are formed they may trigger the generation of the corresponding complex behaviors, which are representations that enclose sequences of actions specified by the behavioral context (see Fig. 10b). In the present work, the behavioral context formation has been oversimplified in order to generate uncomplicated behaviors for the available actuator: the P3DX differential drive. Two basic operations have been defined for the control of the differential drive:

1) ***RotateInPlace***: this operation takes an angle in degrees as input parameter (positive values mean counterclockwise rotation) and triggers the robot rotation in position until it completes the consigned angle.
2) ***MoveStraight***: this operation takes a speed in meters per second as input parameter (positive values mean move forward) and triggers the robot movement towards the current heading (or backwards for negative speed values).

Attending to the relative direction specified by the attention mechanism (a composition of J-Indexes representations), an angle parameter is calculated for the RotateInPlace operation in order to set the robot heading towards the object that "called robot's attention". Also, a speed parameter is calculated as a function of the distance to the object. This means that the typical minimum behavioral context is formed by a sequence of simple actions like a RotateInPlace operation followed by a MoveStraight operation.

IV. Multimodal Integration

Combining multiple monomodal sensory data sources is a typical problem in mobile robotics, also known as multisensory integration or sensor data fusion [16]. Actually, in the present work we are also approaching the problem of fusing proprioceptive and exteroceptive sensor data. Neuroscientists refer to the binding problem [17], as the analogous problem of how to form a unified perception out of the activity of specialized sets of neurons dealing with particular aspects of perception. From the perspective of autonomous robot control we argue that the binding problem can be functionally resolved by applying the proposed contextualization mechanism.

A. Monomodal Context Formation

Taking the bump panel percepts as example, we can illustrate how a sensory context gives place to a monomodal complex percept. Using the aforementioned common criteria, time and location, if the bumper handler of our robot reports contact in bump panels b2, b3, and b4 simultaneously (see Fig. 11), a context is automatically created if these three independent notifications have close enough timestamps. Therefore, the three single percepts are associated by a temporal context. Additionally, as b2, b3, and b4 are located side by side, the corresponding contact percepts J-Indexes will indicate proximity, thus forming an additional spatial context that again associates these three single percepts. The newly created complex percept, which is a composition of three single percepts, also holds a representation of a J-Index. This complex percept J-Index is calculated as a function of the reference vectors of the former single percepts (note that Fig. 11 depicts with solid lines the J-Index referent vectors of the formed complex percept, and dashed lines represent the referent vector of the old single percepts).

The way the J-Index of a complex percept is calculated depends on the nature (shape, dimensions, etc.) of the single percepts that take part in the context that gave place to it. The composition of J-Indexes is trivial when all the single percepts belong to the same modality (as illustrated in Fig. 11). However, the composition can be complex when several different modalities are involved.

B. Multimodal Context Formation

Focusing on the mentioned fundamental criteria for contextualization (time and location), all percepts, independently of their modality, can be compared with each other, thus allowing a simple mechanism to create perceptual contexts. The contexts formed following this method can have significant meaning. For instance, "all objects within the reach of the robot" (context formed applying the criterion of location and estimating that the relative location is below a given threshold, like the robotic arm reach distance in this case), or "all events that took place between five and ten minutes ago" (context formed applying the criterion of time and estimating that the relative timestamp of the events fall within the given

interval). Similarly, more specific criteria can be used in order to build more specific contexts which might not involve all the available sensory modalities. This is the case of the motion and color criteria used in this work.

C. Contextualization Hierarchy

The proposed contextualization mechanism supports hierarchical composition; hence complex percepts can be built by either combining:

- A number of single percepts.
- A number of complex percepts.
- Both single and complex percepts.

In order to assemble coherent percepts, a priority policy has been established in relation to complex percept formation. The first and top priority contextualization step is to build complex percepts that come from the application of contextualization criteria over the same modality single percepts. The outcome of this first step is a set of monomodal complex percepts. As illustrated above, these monomodal complex percepts can come from simultaneous and contiguous bumper contacts or from simultaneous and contiguous salient visual segments. Once the first contextualization step is completed, both the newer monomodal complex percepts and existing single percepts enter the CERA Workspace where multimodal complex percepts are built (see Fig. 13 for an example).

D. Managing contradictory percepts

A common application of multimodal sensory information fusion is the disambiguation or refutation of contradictory sensor data. In the case under study in this paper, contradictory information happen to be processed when the sonar transducers fail to detect a sharp solid corner (the ultrasonic beams are diverted, and do not come back to the transducer, failing to provide a realistic range measurement). In such a scenario, the last resort are the bumpers. When the robot base is too close to the sharp corner, bumpers will contact the obstacle and notify single percepts, which in turn will become complex percepts. However, during the process of complex percepts formation, potential contradictory information has to be handled. The time criteria for context formation will associate the roughly simultaneous readings from both sonar and bumpers. But, in the case of a bad sonar reading the single percepts available are not consistent. Therefore, a policy has to be established in order to build a significant complex percept out of conflicting single percepts. A single but effective approach is to apply a level of confidence to each sensor modality depending on the situation. In the case described here, we have just assigned more confidence to bumper contact notifications than sonar measurements.

V. Paying Attention to Counterpart Robots

Following a counterpart robot across an unknown office-like environment has been selected as a preliminary problem domain for the testing of the proposed attention mechanism. It provides a valid real world scenario where the sensors and actuators described above can be used to achieve the mission goal: to find the *P3DX-Target* robot and then keep a constant distance to it. Both simulated and real environment setups have been prepared. All localization estimation problems have been neglected for the time being. Figure 12 shows a screen capture of the simulated environment we have used for initial testing.

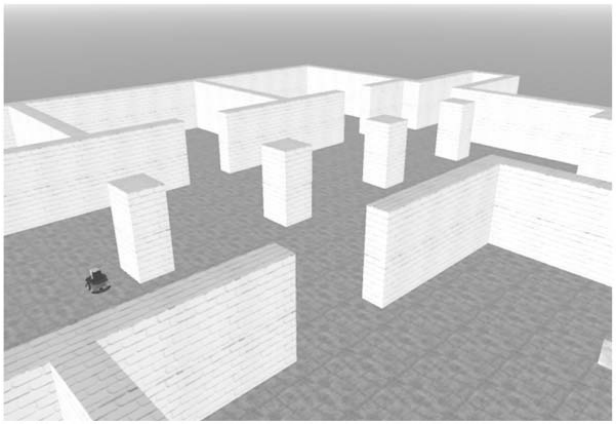

Fig. 12. Simulated indoor environment.

One of the objectives of the proposed attention mechanism is to offer an effective policy for selecting the next action (or complex behavior) as part of the robot's main control loop. In the case of counterpart chasing, spatial contexts are defined in order to estimate the best heading that the robot should take. A specific CERA Instantiation Layer has been coded with the aim of representing the particular complex percepts that are required for the chasing task.

Robot mission is structured in two sequential tasks. Firstly, during the searching task, *P3DX-Chaser* has to find out if the *P3DX-Target* robot is in the surroundings. Secondly, if the target has been detected (target engaged), the chaser has to follow it keeping a constant separation distance. If for some reason the target is loss, *P3DX-Chaser* will come back to the first task. During the initial searching task, the chaser wander randomly performing surrounding inspections (360° turns) periodically. Two attentional contexts are applied during the searching phase in order to detect the presence of the target: "a red and moving object". Detecting this sort of objects involves paying attention to complex percepts that are formed as a result of the conjoined application of color and motion context criteria. Therefore, such a context should be activated during the searching phase. However, as motion criteria is difficult to assess when the own referential system is also moving, CERA Core Layer initially activates only the red color context while the robot is wandering or performing a 360° scan. When a salient complex percept is obtained due to red color saliency (like in Fig. 9), robot comes to a full stop and activates a new attentional context with two criteria: red color and motion. Then, if more complex percepts are obtained from a given location, the target is recognized and engaged, and the second task (following) is activated. During the second task, again a single criterion context for red color is activated and the robot heading is adjusted to the direction indicated by the target complex percepts (SH value and sign). Basically, a j referent vector is calculated based on the target

complex percept J-Index, and simple actions are generated that will cause the chaser to head towards the target.

Keeping a constant distance to the target is facilitated by the ranging data obtained from sonar. As single percepts from vision and single percepts from sonar share location parameters (j referent vectors), distance to target can be estimated by multimodal contextualization. Actually, the complex percepts that represent the target are composed of both visual and sonar single percepts. These single percepts were associated in the same complex percept because of their affinity in relative location. This means that target complex percepts include sonar ranging data in addition to the visual pose estimation.

Fig. 13 shows an example of both visual and sonar data as ingested in the CERA Physical Layer, where associated time and location parameters are calculated. Then, sensor preprocessors build single percepts including timestamps and J-Indexes. All generated single percepts enter the CERA Physical workspace where the complex percepts are built based on current active contexts. Active contexts are established by the higher control system running in the CERA Core Layer. The scenario depicted in figure 13 corresponds to the chasing task; hence a context for red objects is active (in addition to location and time, which are always used as contextualization criteria). Right side of the figure shows an example of the complex percepts that are formed due to the application of the mentioned contextualization mechanism. Single percepts corresponding to visual segments $S_{5,5}$ and $S_{6,5}$ are selected because the present saliency in terms of the red color contextualization criterion. Given that their j referent vectors happen to be contiguous, a new monomodal (visual) complex percept is built as a composition of them. As shown in the picture, the J-Index of the newer monomodal complex percept points to the geometrical center of the visual segment formed as a combination of the two former single percept segments. It can be noticed that the J-Index of this visual complex percept does not spot the actual center of the target, but the approximation is good enough for the realtime chasing task. Once monomodal complex percepts have been built, time and location contextualization is applied amongst different modalities.

Right bottom representation in the picture (Fig. 13) corresponds to sonar j referent vectors, including a highlighted single percept (the one obtained from the reading of the sonar transducer oriented at $+10°$). The projection outlined top-down from the visual complex percept to this sonar single percept indicates that both percepts are to be associated and will form a multimodal complex percept. Time association is obvious; however, the location contextualization between visual and sonar percepts require some additional parametric alignment as these different modality sensors present particular orientations and wide span. Furthermore, as explained above, only the X coordinate is considered for visual percepts (visual horizontal axis). While we have used a $+90°$ field of view camera, the Pioneer 3DX robot frontal sonar ring covers a total field of $+195°$ (including blind angles between transducer cones). Therefore, only percepts originated from the central $90°$ of sonar coverage are taken into account for visual to sonar contextualization. Black dashed lines on the right hand side of the figure represent the alignment between visual horizontal axis and the central $-45°$ to $+45°$ angular span of frontal sonar. In this case, the value of SH in the visual complex percept corresponds to the sonar percept originated from the sonar transducer at $+10°$. The measurement represented in this particular sonar percept (2493 millimeters) is directly the distance estimate assigned to the multimodal complex percept as visual percept itself does not provide any distance estimate.

Preliminary results obtained applying the proposed attention mechanism to the human-controlled target chasing task are promising; however more complex environments have to be tested in order to appreciate the real potential of the cognitive approach. In addition to the manual control of *P3DX-Target*, which produces very variable results, three autonomous simple behaviors have been implemented with the aim to test the capability of the attention mechanism when confronted to different movement patterns (scenarios a, b and c depicted in Fig. 14). Fig. 14 shows the typical trajectories of the autonomous control schemes implemented in *P3DX-Target*. Initial time to target engaged state varies and is basically dependent on start position of both *P3DX-Chaser* and *P3DX-Target* robots. Therefore, in the present case, the performance of the attention mechanism is measured in terms of the overall duration of target engaged state (percentage of total navigation time when the target is engaged). The performance when chasing targets in open space (wide corridors) is 100% in scenario (a), and close to 100% in scenarios (b) and (c). However, when the target (a, b, or c) performs obstacle avoidance maneuvers close to walls, performance usually fall to 50-70%. In these situations the chaser also has to avoid obstacles, eventually causing the loss of target.

VI. Conclusion and Future Work

A novel attention mechanism for autonomous robots has been proposed and preliminary testing has been done in the domain of simple mobile object recognition and chasing. The integration of the attention cognitive function into a layered control architecture has been demonstrated. Additionally, the problem of multimodal sensory information fusion has been addressed in the proposed approach using a generic context formation mechanism. Preliminary results obtained with the simulator show that this account is applicable to classical mobile robotics problems. Nevertheless, moving to a real world environment and facing more demanding missions including localization requirements would imply dealing with the problem of imperfect odometry [18]. In such a scenario our proposed attention mechanism had to be integrated into a SLAM (Simultaneous Localization and Mapping) system.

The attention mechanism proposed in this work is designed to be highly dynamic and configurable. Following the same principles described above, more contexts can be created as more contextualization criteria are defined in the system. The concrete definition of criteria and context is to be selected based on the specific problem domain.

The system described in this paper is work in progress. Counterpart recognition is currently based on color and movement detection, however we are working on adding other

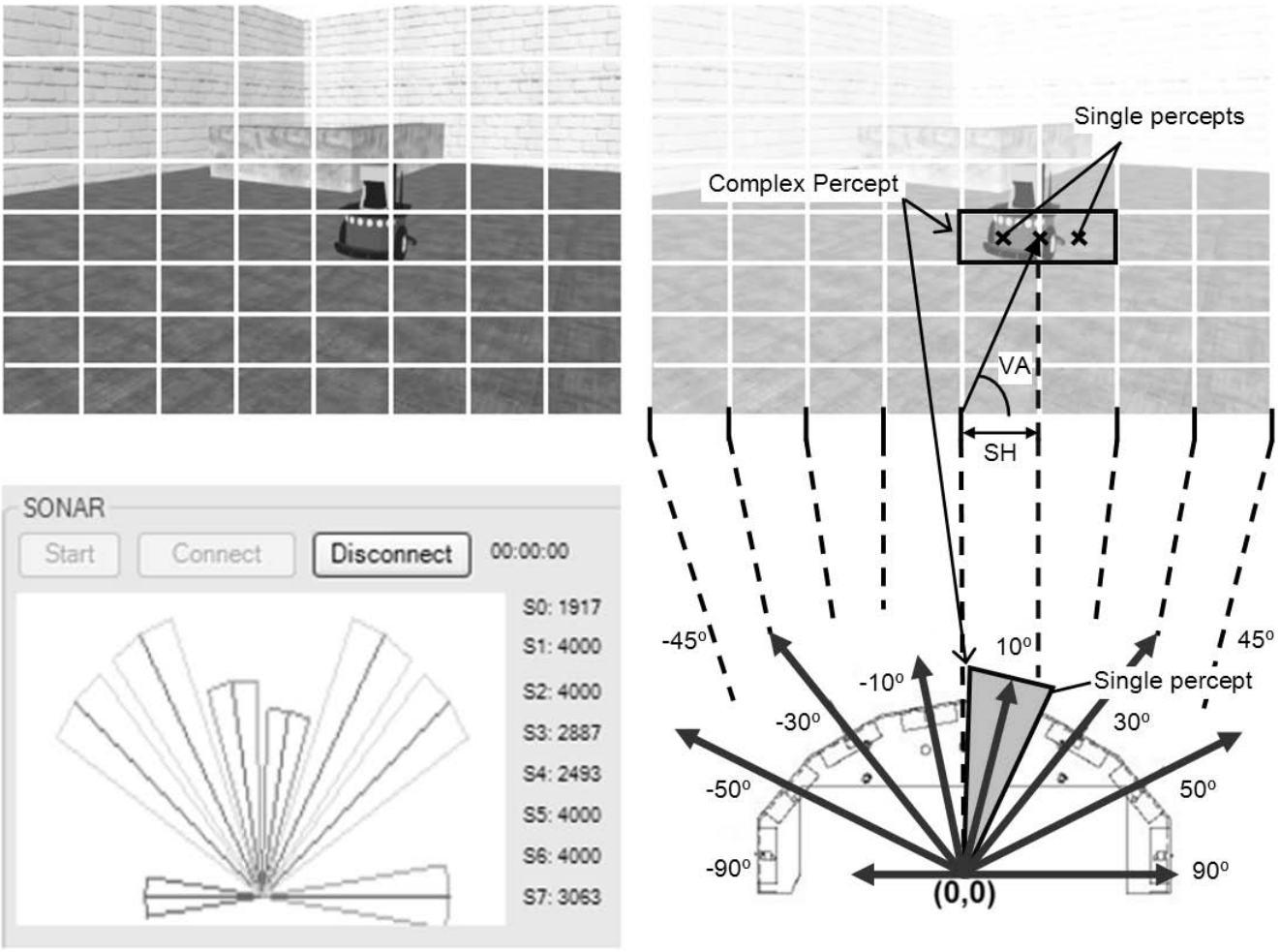

Fig. 13. Upper left image is a representation of the segmented visual sensory information acquired by the simulated onboard camera. Lower left graph is a capture from the CERA graphical user interface displaying real time sonar transducer measurements. Sonar ranging representation capture corresponds to the particular instant when the camera acquired the image depicted above. Right hand side of the picture shows a representation of the multimodal complex percept being built with this sensory information

visual and higher cognitive properties recognition in order to build a most robust mechanism. Concretely, visual texture recognition, vertical symmetry detection, and movement pattern identification are expected to greatly improve robustness in real world environments. Furthermore, a mechanism for the detection of structurally coherent visual information could be implemented as part of the proposed attention mechanism. Complex percepts formed representing unique objects could be evaluated in terms of their structural coherence, as human brain seems to do [19]. More complex attentional contexts (and therefore more contextual criteria) have to be defined in order to face other problem domains. Perception is well covered for sonar range finder and bumpers. However, additional development is required in the CERA Physical Layer in order to add more functionality to visual perception, e.g. visual distance estimation. The definition of behavioral contexts and complex behaviors should also be enhanced to cope with more complex actuators and to generate more efficient behaviors. At the level of the CERA Core Layer, learning mechanisms could be applied in order to improve the attention selection technique. Moreover, the attention mechanism is to be integrated with other Core Layer modules, like memory and self-coordination modules in order to use the required related information for the activation of appropriate contexts in the Instantiation and Physical layers.

Given the need to define complex spatiotemporal relations in the process of attentional contexts formation, the application of fuzzy temporal rules will be considered as they have been proved to be an effective method in landmark detection (like doors) for mobile robots [20].

ACKNOWLEDGMENT

This research work has been supported by the Spanish Ministry of Education and Science CICYT under grant TRA2007-67374-C02-02.

REFERENCES

[1] P. J. Burt, “Attention mechanisms for vision in a dynamic world,” pp. 977–987 vol.2, 1988.

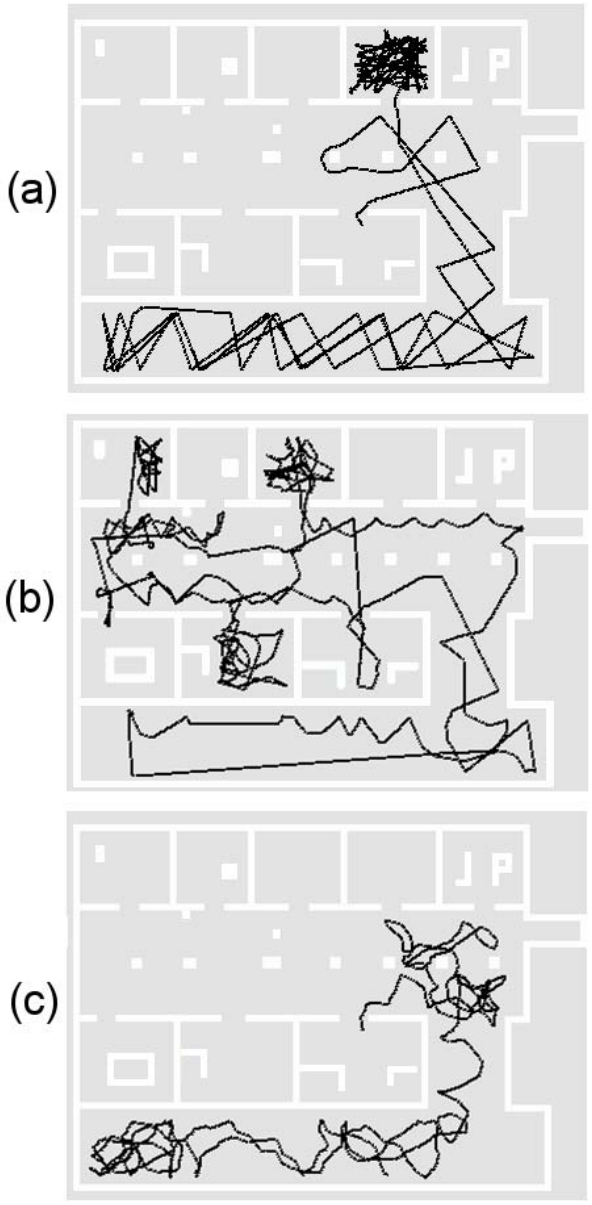

Fig. 14. Behavior (a) is the outcome of the simplest control algorithm which is based on performing random turns when the target is too close to an obstacle. Behavior (b) is obtained by adding an attentional bias to unvisited areas. Finally, behavior (c) adds random turns to the former control strategies. These three simple autonomous control strategies have been used to implement a moving target in the Robotics Developer Studio simulation environment.

[2] R. Arrabales and A. Sanchis, "Applying machine consciousness models in autonomous situated agents," *Pattern Recognition Letters*, vol. 29, no. 8, pp. 1033–1038, 6/1 2008.

[3] J. L. C. Mariño, "Una arquitectura de atención distribuida para agentes con sensorizacin multimodal", Doctoral Thesis. Universidade de Coruña, 2007.

[4] S. Frintrop, A. Nchter, and H. Surmann, "Visual attention for object recognition in spatial 3d data," in *2nd International Workshop on Attention and Performance in Computational Vision*, ser. Lecture Notes in Computer Science, vol. 3368. Springer, 2004, pp. 168–182.

[5] R. Arrabales, A. Ledezma, and A. Sanchis, "Modeling consciousness for autonomous robot exploration," in *IWINAC 2007*, ser. Lecture Notes in Computer Science, vol. 4527-4528, 2007.

[6] B. J. Baars, *A Cognitive Theory of Consciousness*. New York: Cambridge University Press, 1993.

[7] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, 1995.

[8] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. Bradski, P. Baunstarck, S. Chung, and A. Y. Ng, "Peripheral-foveal vision for real-time object recognition and tracking in video", in *Proceeding of the International Joint Conference on Artificial Intelligence*, 2007, pp. 2115–2121.

[9] M. H. Giard and F. Peronnet, "Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study", *The Journal of Cognitive Neuroscience*, vol. 11, no. 5, pp. 473–490, September 1 1999.

[10] D. Senkowski, D. Talsma, M. Grigutsch, C. S. Herrmann, and M. G. Woldorff, "Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations", *Neuropsychologia,*, vol. 45, no. 3, pp. 561–571, 2007.

[11] C. Spence and S. Squire, "Multisensory integration: Maintaining the perception of synchrony", *Current Biology*, vol. 13, no. 13, pp. R519–R521, 7/1 2003.

[12] L. Fogassi, V. Gallese, G. Pellegrino, L. Fadiga, M. Gentilucci, G. Luppino, M. Matelli, A. Pedotti, and G. Rizzolatti, "Space coding by premotor cortex", *Experimental Brain Research*, vol. 89, no. 3, pp. 686–690, 06/01 1992.

[13] M. Avillac, S. Deneve, E. Olivier, A. Pouget, and J.-R. Duhamel, "Reference frames for representing visual and tactile locations in parietal cortex", *Nature neuroscience*, vol. 8, no. 7, pp. 941–949, 2005.

[14] S. Zeki, J. Watson, C. Lueck, K. Friston, C. Kennard, and R. Frackowiak, "A direct demonstration of functional specialization in human visual cortex", *Journal of Neuroscience*, vol. 11, no. 3, pp. 641–649, March 1 1991.

[15] I. Aleksander and B. Dunmall, "Axioms and tests for the presence of minimal consciousness in agents", *Journal of Consciousness Studies*, vol. 10, no. 4-5, 2003.

[16] R. C. Luo, "Multisensor integration and fusion in intelligent systems", pp. 901–931, 1989.

[17] A. Revonsuo and J. Newman, "Binding and consciousness", *Consciousness and Cognition*, vol. 8, no. 2, pp. 123–127, 6 1999.

[18] S. Thrun, "Probabilistic algorithms in robotics", *AI Magazine*, vol. 21, no. 4, pp. 93–109, 2000.

[19] D. L. Schacter, E. Reiman, A. Uecker, M. R. Roister, L. S. Yun, and L. A. Cooper, "Brain regions associated with retrieval of structurally coherent visual information," *Nature*, vol. 376, pp. 587–590, aug 1995.

[20] P. Carinena, C. V. Regueiro, A. Otero, A. J. Bugarin, and S. Barro, "Landmark detection in mobile robotics using fuzzy temporal rules", pp. 423–435, 2004.