Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2011, Article ID 841078, 10 pages
doi:10.1155/2011/841078

Research Article

A Novel Biologically Inspired Attention Mechanism for a Social Robot

Antonio Jesús Palomino, Rebeca Marfil, Juan Pedro Bandera, and Antonio Bandera

Grupo ISIS, Departamento de Tecnología Electrónica, E.T.S.I. Telecomunicación, Universidad de Málaga, Campus de Teatinos, 29071 Málaga, Spain

Correspondence should be addressed to Antonio Bandera, ajbandera@uma.es

Received 16 June 2010; Revised 8 October 2010; Accepted 19 November 2010

Academic Editor: Steven McLaughlin

In biological vision systems, the attention mechanism is responsible for selecting the relevant information from the sensed field of view. In robotics, this ability is specially useful because of the restrictions in computational resources which are necessary to simultaneously perform different tasks. An emerging area in robotics is developing social robots which are capable to navigate and to interact with humans and with their environment by perceiving the real world in a similar way that people do. In this proposal, we focus on the development of an object-based attention mechanism for a social robot. It consists of three main modules. The first one (preattentive stage) implements a concept of saliency based on "proto-objects." In the second stage (semiattentive), significant items according to the tasks to accomplish are identified and tracked. Finally, the attentive stage fixes the field of attention to the most salient object depending on the current task.

1. Introduction

In the last few years, emphasis has increased in the development of robot vision systems according to the model of natural vision due to its robustness and adaptability. Research in psychology and physiology demonstrates that the efficiency of natural vision has foundations in visual attention, which is a process that filters out irrelevant information and limits processing to items that are relevant to the present task [1]. Developing computational perception systems that provide these same sorts of abilities is a critical step in designing social robots that are able to cooperate with people as capable partners, that are able to learn from natural human instruction and that are intuitive and engaging for people to interact with, but that are also able to simultaneously navigate in initially unknown environments or to perform other tasks as, for instance, grasping a specific object.

In the literature, methods to model attention are categorized in space-based and object-based. The fundamental difference between them is the underlying unit of attentional selection [2]. While space-based methods deploy attention at the level of space locations, the object-based theory holds that a preattentive process segments the image into objects and then the attention is allocated to these objects. The models of space-based attention scan the scene by shifting attention from one location to the next to limit the processing to a variable size of space in the visual field. Therefore, they have some intrinsic disadvantages. In a normal scene, objects may overlap or share some common properties. Then, attention may need to work in several discontinuous spatial regions at the same time. On the other hand, if different visual features, which constitute the same object, come from the same region of space, an attention shift will be not required [3]. Object-based models of visual attention provide a more efficient visual search than space-based attention. Besides, it is less likely to select an empty location. In the last few years, these models of visual attention have received an increasing interest in computational neuroscience and in computer vision. Object-based attention theories are based on the assumption that attention must be directed to an object or group of objects, instead of to a generic region of the space [4]. In fact, neurophysiological studies [2] show that, in selective

attention, the boundaries of segmented objects, and not just spatial position, determine what is selected and how attention is deployed. Therefore, these models will reflect the fact that the perception abilities must be optimized to interact with objects and not just with disembodied spatial locations. Thus, visual systems will segment complex scenes into objects which can be subsequently used for recognition and action. However, recent psychological research shows that, in natural vision, the preattentive process divides a visual input into raw or primitive objects [5] instead of well-defined objects. Some authors use the notion of *proto-objects* [4, 6] to refer to these primitive objects, that are defined as units of visual information that can be bound into a coherent and stable object. On the other hand, other challenging issue in visual attention models is the *inhibition of return*. This process avoids continuous attention to only one location or object. The most used approach is to build an inhibition map that contains suppression factors for previously attended regions [7, 8]. The problem of these maps is that they are not able to manage inhibited moving objects or situations where the vision system is moving. To deal with these situations, it is necessary to track the inhibited objects [4, 9].

Following these considerations, this paper presents a general object-based visual attention model which exploits the concept of proto-objects as image entities which do not necessarily correspond with a recognizable object, although they possess some of the characteristics of objects [4, 10]. Thus, it can be considered that they are the result of the initial segmentation of the image input into candidate objects (i.e., grouping together those input pixels which are likely to correspond to parts of the same object in the real world, separately from those which are likely to belong to other objects). This is the main contribution of the proposed approach, as it is able to group the image pixels into entities which can be considered as *segmented perceptual units* using a novel perceptual segmentation algorithm in a preattentive stage. Once the input image has been split, the saliency of each region is evaluated by combining four low-level features. In this combination process, the weight of each evaluated feature will depend on the performed task. Other important contribution is the inclusion of a semiattentive stage which will take into account the currently executed tasks in the information selection process. Besides, it is capable of handling dynamic environments where the locations and shapes of the objects may change due to motion and minor illumination differences between consecutive acquired images. In order to deal with these scenes, a mean shift-based tracking approach [11] for inhibition of return is employed. Recently attended proto-objects will be stored in a memory module for several fixations. Thus, if the task requires to shift the focus of attention to a previously attended proto-object and it is still stored in this memory, these fixations could be fastly executed. Finally, an attentive stage is included where two different behaviors or tasks have been programmed. Currently, these behaviors only need visual information to be accomplished and thus they will allow to test the performance of the proposed visual perception system.

The remainder of the paper is organized as follows. Section 2 provides a brief related work. Section 3 presents an overview of the proposed attention model. The preattentive, semiattentive, and attentive stages of the proposal are described in Sections 4, 5, and 6, respectively. Section 7 deals with some obtained experimental results. Finally, conclusions are shown in Section 8.

2. Related Work

There are mainly two psychological theories of visual attention that have influenced the computation models existing today [12]: the feature integration theory and the guided search. The feature integration theory proposed by Treisman and Gelade [13] suggests that the human vision system detects separable features in parallel in an early step of the attention process. Then, they are spatially combined to finally attend individually to each relevant location. According to this model, methods compute image features in a number of parallel channels in a preattentive task-independent stage. The extracted features are integrated into a single saliency map which codes the saliency of each image pixel [12, 14–16]. While this previous theory is mainly based on a bottom-up component of attention, the guided search theory proposed by Wolfe et al. [17, 18] is centered on the fact that a top-down component in attention can increase the speed of the process when identifying the presence of a target in a scene. The model computes a set of features over the image and the top-down component activates locations that might contain the features of the searched target. These two approaches are not mutually exclusive, and nowadays, some efforts in computational attention are being conducted to develop models which combine a bottom-up preattentive stage with a top-down attentive stage [19]. The idea is that while the bottom-up step is independent of the task, the top-down component tries to model the influence of the current executed task in the process of attention. Therefore, Navalpakkam and Itti [19] extended Itti's model [14] by building a multiscale object representation in a long-term memory. The multiscale object features stored in this memory determine the relevance of the scene features depending on the current executed task.

The aforementioned computational models are space-based methods which allocate the attention to a region of the scene rather than to an object or proto-object. An alternative to space-based methods was proposed by Sun and Fisher in [3]. They present a grouping-based saliency method and a hierarchical selection of attention at different perceptual levels (points, regions, or objects). The problem of this model is that the groups are manually drawn. Orabona et al. [4] propose a model of visual attention based on the concept of "proto-objects" as units of visual information that can be bound into a coherent and stable object. They compute these proto-objects by employing the watershed transform to segment the input image using edge and colour features in a preattentive stage. The saliency of each proto-object is computed taking into account top-down information about the object to search depending on the task. Yu et al. [6] propose a model of attention in which, first in a preattentive stage the scene is segmented into "proto-objects"

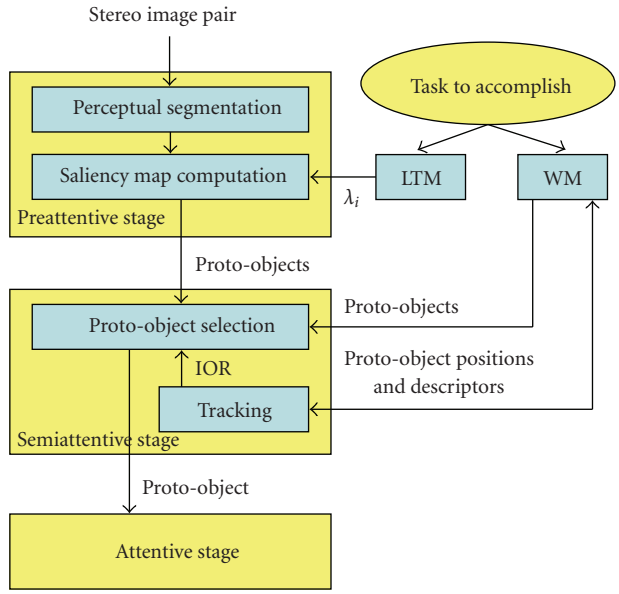


Figure 1: Overview of the proposed model of visual attention.

in a bottom-up manner using Gestalt theories. After that, in a top-down way, the saliency of the proto-objects is computed taking into account the current task to accomplish by using models of objects which are relevant to this task. These models are stored in a long-term memory.

3. Overview of the Proposed Model of Attention

This paper presents an object-based model of visual attention for a social robot which works in a dynamic scenario. The proposed system integrates task-independent bottom-up processing and task-dependent top-down processing. The bottom-up component determines the set of proto-objects present in the image. It also describes them by a set of low-level features that are considered relevant to determine their corresponding saliency values. On the other hand, the top-down component weights the low-level features which characterize each proto-object to obtain a single saliency value depending on the task. From the recently attended proto-objects, it also selects those which are relevant for the task.

Figure 1 shows an overview of the proposed architecture. The visual attention model implements a concept of salience based on proto-objects which are computed in the preattentive stage of this model. These proto-objects are defined as the blobs of uniform colour and disparity of the image which are bounded by the edges obtained using a Canny detector. A stereo camera is used to compute a dense disparity map. At the pre-attentive stage, proto-objects are described by four low-level features, which are computed in a task-independent way: colour and luminosity contrasts between the proto-object and all the objects in its surroundings, mean disparity, and the probability of the "proto-object" to be a face or a hand taking into account its colour. A proto-object catches the attention if it differs from its immediate surroundings or if its associated low-level features are interesting for the task to reach. A weighted normalized summation is employed to combine these features into a single saliency map. Depending on the current task to perform, different sets of weights are chosen. These task-dependent weights will be stored in a memory module. In our proposal, this module will be called the long-term memory (LTM), as it resembles the one proposed by Borji et al. [20]. The main steps of the pre-attentive stage of the proposed attention mechanism are resumed in Algorithm 1. This pre-attentive stage is followed by a semiattentive stage where a tracking process is performed over the recently attended proto-objects using a mean shift-based algorithm [11]. The output regions of the tracking algorithm are used to implement the inhibition of return (IOR). This stage is resumed in Algorithm 2. The IOR will avoid revisiting recently attended objects. To store these attended proto-objects, we include at this level a working memory (WM) module. This module has a fixed size, and stored patterns should be forgotten after several fixations to include new proto-objects. It must be noted that our two proposed memory modules are not exactly related to the memory organization postulated by the cognitive psychology or neuroscience. They satisfy specific functions in the proposed architecture.

Algorithm 1 (Pre-attentive stage). We have the following

(1) Pre-segmentation of the input image into homogeneous colour blobs
(2) Perceptual grouping of the blobs into proto-objects
(3) Computation of the features associated to each proto-object: colour contrast, intensity contrast, disparity and skin colour
(4) Computation of attractivity maps for each of the computed features
(5) Combination of the attractivity maps into a final saliency map (Ec. (1))

end

Algorithm 2 (Semiattentive stage). We have the following

(1) Tracking of the most salient proto-objects which has been already attended and which are stored in the WM

(2) IOR over the saliency map SRTATE selection of the most salient proto-objects of the current frame

(3) Updating of the WM

end

When a new task has to be performed by the robot, the system looks in the WM for the proto-object (or proto-objects) which is necessary to accomplish the task. If the proto-object has not been recently attended, the system looks in the LTM for the best set of weights for obtaining the saliency map according to the task to reach. Then, pre-attentive and semiattentive stages are performed. On the other hand, if the proto-object required by the task is stored in the WM, then it will be possible to recover its position in the scene from the WM and to send this data to the attentive stage. In this case, the pre-attentive and semiattentive stages are also performed, but now using a set of weights which does not enhance any specific feature in the saliency map computation (generic exploration behaviour). If new proto-objects are now found, they could launch a different task. In any case, it must be noted that to solve the action-perception loop is not the goal of this work, which is focused on the visual perception system.

Finally, in order to test the proposed perception system, we have developed two specific behaviours. The human gesture recognition module and the visual landmark detector are the responsible for recognize the upper-body gestures of a person who is interacting with the robot and to provide visual natural landmarks for mobile robot navigation, respectively. They will be further described in Section 6.

4. Preattentive Stage: Object-Based Selection

As it was aforementioned in Section 1, several psychological studies have shown that, in natural vision, the visual input is divided into proto-objects in a preattentive process [5]. Following this guideline, the proposed model of attention implements a pre-attentive stage where the input image is segmented into perceptually uniform blobs or proto-objects. In our case, these proto-objects are defined as the union of a set of blobs of uniform colour and disparity of the image which will be partially or totally bounded by the edges obtained using a Canny detector. As the process to group image pixels into higher-level structures can be computationally complex, perceptual segmentation approaches typically combine a presegmentation step with a subsequent perceptual grouping step [21]. The pre-segmentation step performs the low-level definition of segmentation as the process of grouping pixels into homogeneous clusters, and the perceptual grouping step conducts a domain-independent grouping which is mainly based on properties such as the proximity, closure, or continuity.

In our proposal, both steps are performed using an irregular pyramid: the Bounded Irregular Pyramid (BIP) [22]. Pyramids are hierarchical structures which have been widely used in segmentation tasks [22]. Instead of performing image segmentation based on a single representation of the input image, a pyramid segmentation algorithm describes the contents of the image using multiple representations with decreasing resolution. Pyramid segmentation algorithms exhibit interesting properties when compared to segmentation algorithms based on a single representation. Thus, local operations can adapt the pyramid hierarchy to the topology of the image, allowing the detection of global features of interest and representing them at low resolution levels [23]. With respect to other irregular pyramids, the main advantage of the BIP is that it is able to obtain similar segmentation results but in a faster way [22, 24]. Hence, the proposed approach uses the BIP to accomplish the detection of the proto-objects. In this hierarchy, the first levels perform the pre-segmentation step using a colour-based distance to group pixels into homogeneous blobs (see [22, 25] for further details). After this step, grouping blobs aims at simplifying the content of the obtained image partition in order to extract the set of final proto-objects. For managing this grouping, the BIP structure is also used: the obtained pre-segmented blobs constitute the first level of the perceptual grouping hierarchy, and successive levels are built using a distance which integrates edge and region descriptors [21]. Figure 2 shows a pre-segmentation image and the final regions obtained after applying the perceptual grouping. It can be noted that the pre-segmentation approach has problems to merge regions in shaded tones (e.g., wall left part). Although the perceptual grouping step solves some of these problems, the final regions obtained by the described bottom-up process may not always correspond to the natural image objects.

Once the set of proto-objects has been obtained, the saliency of each of them is computed and stored in a saliency map. To do that, four features are computed for each proto-object i: colour contrast (MCG_i), intensity contrast (MLG_i), disparity (D_i), and skin colour (SK_i). From these four features, attractivity maps are computed, containing high values for interesting proto-objects and lower values for other regions in a range of $[0 \cdots 255]$. Finally, similarly to other models [9, 26], the saliency map is computed by combining the feature maps into a single representation. A weighted normalized summation has been used as feature combination strategy because, although this is the worst strategy when there are a big number of feature maps [27], it has been demonstrated that its performance is good in systems with a small number of feature maps. Then, the final saliency value, Sal_i of each proto-object, i, is computed as

$$\mathrm{Sal}_i = \lambda_1 \mathrm{MCG}_i + \lambda_2 \mathrm{MLG}_i + \lambda_3 D_i + \lambda_4 \mathrm{SK}_i \quad (1)$$

being $\{\lambda\}_{i=1\cdots4}$ the weights associated to each feature map which values are set depending on the current task to execute in the attentive stage. These λ_i values are stored in the LTM. In our current implementation, only two different behaviours can be chosen at the attentive stage.

(a) (b) (c)

Figure 2: Pre-attentive stage: (a) original left image; (b) pre-segmentation image; and (c) final set of proto-objects.

The first one looks for visual landmarks for mobile robot navigation giving, more importance to the colour and intensity contrasts ($\lambda_1 = \lambda_2 = 0.35$ and $\lambda_3 = \lambda_4 = 0.15$), and the second one looks for humans to interact, giving more importance to the skin colour map ($\lambda_1 = \lambda_2 = 0.15$, $\lambda_3 = 0.30$, and $\lambda_4 = 0.40$). In any case, the setting of these parameters must be changed in future versions, including a reinforcement learning approach which allows to choose these values from different trials in an unsupervised manner.

5. Semiattentive Stage: the Inhibition of Return and the Role of the Working Memory

Psychophysics studies about human visual attention have established that a local inhibition is activated in the saliency map when a proto-object is already attended. This mechanism avoids directing the focus of attention to a proto-object immediately visited, and it is usually called *inhibition of return* (IOR). In the context of artificial models of visual attention, the IOR is typically implemented using a 2D inhibition map which contains suppression factors for one or more focuses of attention recently attended. This approach is valid to manage static scenarios, but it is not able to handle dynamic environments where inhibited proto-objects or the vision itself are in motion, or when minor illumination differences between consecutive frames cause shape changes in the proto-objects. In these scenarios, it is necessary to match proto-objects among consecutive video frames and to move the suppression factors.

Some proposed models, like the Backer et al.'s approach [28], try to solve this problem relating the inhibition to features of activity clusters. However, the scope of dynamic inhibition becomes very limited because it is not related to objects. Thus, we propose an object-based IOR which is implemented using an object tracking procedure. Specifically, the IOR has been implemented using a tracker based on the Dorin Comaniciu's *meanshift* approach [11]. Thus, our approach keeps on tracking the proto-objects that have been already attended in previous frames and which are stored in the WM. Once the new positions of the attended proto-objects are obtained, a suppression mask image is generated and the regions of the image which are associated to already attended proto-objects are inhibited in the current saliency map (i.e., these regions have a null saliency value).

As it has been aforementioned, the working memory (WM) has an important role in the top-down part of the proposed system as well as to address the inhibition of return. Basically, this memory module is the responsible for storing the recently attended proto-objects. To do that, a set of descriptors of each proto-object is stored, its colour histogram regularized by a spatial kernel (required by the mean-shift algorithm), its mean colour (obtained in the perceptual grouping step), its pre-attentive features (colour and intensity contrasts, mean disparity and skin colour), its position in the scene, and its time to live. It must be noted that the proposed pre-attentive and semiattentive stages have been designed as early visual processes. That is, object recognition cannot be performed at these stages because it is considered a more complex task, that will be carried out in later stages of the visual process. For this reason, the search of a proto-object required by the task in the WM is only accomplished based on its mean colour and on its associated pre-attentive features. This set of five features will be compared with those stored in the WM using a simple Euclidean distance. The time to live determines when a stored pattern should be removed from the WM.

6. Attentive Stage

A social robot is a robot that must be capable to interact with its environment and with humans and other robots. Among the large set of behaviours that this kind of robots must exhibit, we have implemented two basic behaviors in this stage: a visual natural landmark detector and a human gesture recognition behavior. These behaviors are the responsible for provide natural landmarks for robot navigation and to recognize the upper-body gestures of a person which is interacting with the robot, respectively. It is clear that the robot would need other behaviors to develop its activities in a dynamic environment (e.g., to solve path planning and obstacle avoidance tasks or to exhibit verbal human-robot interaction abilities). However, these two implemented behaviors will allow to test the capacity of the pre-attentive (and semiattentive) stages to provide good candidate proto-objects to higher-level modules of attention.

Specifically, among the set of proto-objects, the visual landmark detector task should select the set of them which

are very contrasted in colour with their surroundings. In order to do that, the weights used in the saliency computation give more importance to the colour and intensity contrasts maps over the rest ones (as it was previously mentioned in Section 4). Among the set of more salient proto-objects in the final saliency map, the visual landmark detection behaviour chooses those which satisfy certain conditions. The key idea is to use as landmarks quasi-rectangular-shaped proto-objects without significant internal holes and with a high value of saliency. In this way, we try to avoid the selection of segmentation artifacts, assuming that a rectangular region has less probability to be a segmentation error than a sparse region with a complex shape. Selected proto-objects cannot be located at the image border in order to avoid errors due to partial occlusions. On the other hand, in order to assure that the regions are almost planar, regions which present abrupt depth changes inside them are also discarded. Besides, it is assumed that large regions could be more likely associated to nonplanar surfaces. Finally, the selection of proto-objects with a high value of saliency guarantees a higher probability of repeatability than non-salient ones. A detailed explanation of this behavior can be found in [24].

On the other hand, social robots are robots that are not only aware of their surroundings. They are also able to learn from, recognize, and communicate with other individuals. While other strategies are possible, robot learning by imitation (RLbI) represents a powerful, natural, and intuitive mechanism to teach social robots new tasks. In RLbI scenarios, a person can teach a robot by simply demonstrating the task that the robot has to perform. The behaviour included in the attentive stage of the proposed attention model is an RLbI architecture that provides a social robot with the ability to learn and to imitate upper-body social gestures. A detailed explanation of this architecture can be found in Bandera et al. [29]. The inputs of the architecture are the face and the hands of the human demonstrator and her silhouette. The face and the hands are obtained using the face detector proposed by Viola and Jones [30], which is executed over the most salient skin coloured proto-objects obtained in the semiattentive stage. In order to obtain this proto-objects, the weights to compute the final saliency map give more importance to the skin colour feature map (as it was mentioned in Section 4).

7. Results

Different tests have been performed to evaluate the ability of the proposed detector to extract salient regions, the stability of these regions, and the capacity of the tracking algorithm to correctly implement the dynamic inhibition of return. With respect to the attention stages, we have also tested the ability of this attention mechanism to provide visual landmarks for environment mapping in a mobile robots navigation framework and to provide skin-coloured regions to a human gesture recognition system. In these two application areas, the proposed visual perception system was tested using a stereo head mounted on a mobile robot. This robot, named NOMADA, is a new 1.60 meters tall robot that is currently being developed in our research group. It has wheels for holonomic movements and is equipped with different types of sensors, an embedded PC for autonomous navigation, and a stereo vision system. The current mounted stereo head is the STH-MDCS from Videre Design, a compact, low-power colour digital stereo head with an IEEE 1394 digital interface. It consists of two 1.3 megapixel, progressive scan CMOS imagers mounted in a rigid body, and a 1394 peripheral interface module, joined in an integral unit. Images are restricted to 640×480 or 320×240 pixels. The embedded PC, that processes these images using the Linux operating system, is a Core 2 Duo at 2.4 Ghz, equipped with 1 Gb of DDR2 memory at 800 Mhz and 4 Mb of cache memory.

7.1. Evaluating the Performance of the Proposed Salient Region Detector. The proposed model of visual attention has been qualitatively examined through video sequences which include humans and other moving objects in the scene. Figure 3 shows the left images of several image pairs of an image sequence perceived from a stationary binocular camera head. Although the index values below each image are not consecutive, all image pairs are processed. The attended proto-object is marked by a red bounding-box in the input frames. Proto-objects which are inhibited are marked by a white bounding-box. Only one proto-object is attended at each fixation. Among the inhibited proto-objects, there are static items, such as the blue battery attended in frame 10, but also dynamic ones, such as the hands attended in frames 20 or 45.

The inhibition of static proto-objects will be discarded when they remain in the WM for more than a specific number of frames (specified by their time to live). That is, when the time to live of a proto-object expires, it is removed from the WM; thus, it could be attended again (e.g., the blue battery enclosed by the focus of attention at frames 10 and 55). Additionally, the inhibition of dynamic proto-objects will be also discarded if the tracking algorithm detects that they have suffered a high shape deformation (this is the reason for discarding the inhibition of the right hand after frame 50) or when they disappear from the field of view (e.g., the blue cup after frame 15). On the other hand, it must be noted that the tracker follows the activity of the inhibited proto-objects very closely, preventing the templates employed by the mean-shift algorithm to be corrupted by occlusions. In our case, the tracker is capable of handling scale changes, object deformations, partial occlusions, and changes of illumination. Finally, it can be noted that the focus of attention is directed at certain frames to uninterested regions of the scene. For instance, this phenomenon occurs at frames 15, 40, or 50. It is usual that these regions will be associated to segmentation artifacts. In this case, they may not be correctly tracked because their shapes change excessively over time. As, it has been aforementioned, they are removed from the list of proto-objects stored at the WM.

Figure 3: Left input images of a video sequence. Attended proto-objects have been marked by red bounding-boxes and inhibited ones have been marked by white bounding-boxes.

7.2. Testing the Approach in a Visual Landmarks Detection Framework. To test the validity of the proposed approach to detect stable visual landmark, data were collected driving the robot through different environments while capturing real-life stereo images. Figures 4(a)–4(c) show the results associated to several video frames obtained from three different trials. Visual landmarks are matched using the descriptor and scheme proposed in [24]. Represented proto-objects have been stored in the WM when they were attended and tracked between subsequently acquired frames. In the illustrated frames, the robot is in motion, so all detected visual landmarks are dynamic. As it has been aforementioned, they will be forgotten after several fixations or when they disappear from the field of view. The indexes marked on the figure can be only employed to identify what landmarks have been matched in each video sequence. Thus, they are not a valid reference to match landmarks among the three illustrated sequences. Unlike other methods, such as the Harris-Affine and Hessian-Affine [31] techniques, this approach does not rely on the extraction of interest point features or on differential methods in a preliminary step. It thus provides complementary image information, being more closely related to those region detectors based on image intensity analysis, such as the MSER and IBR approaches [31].

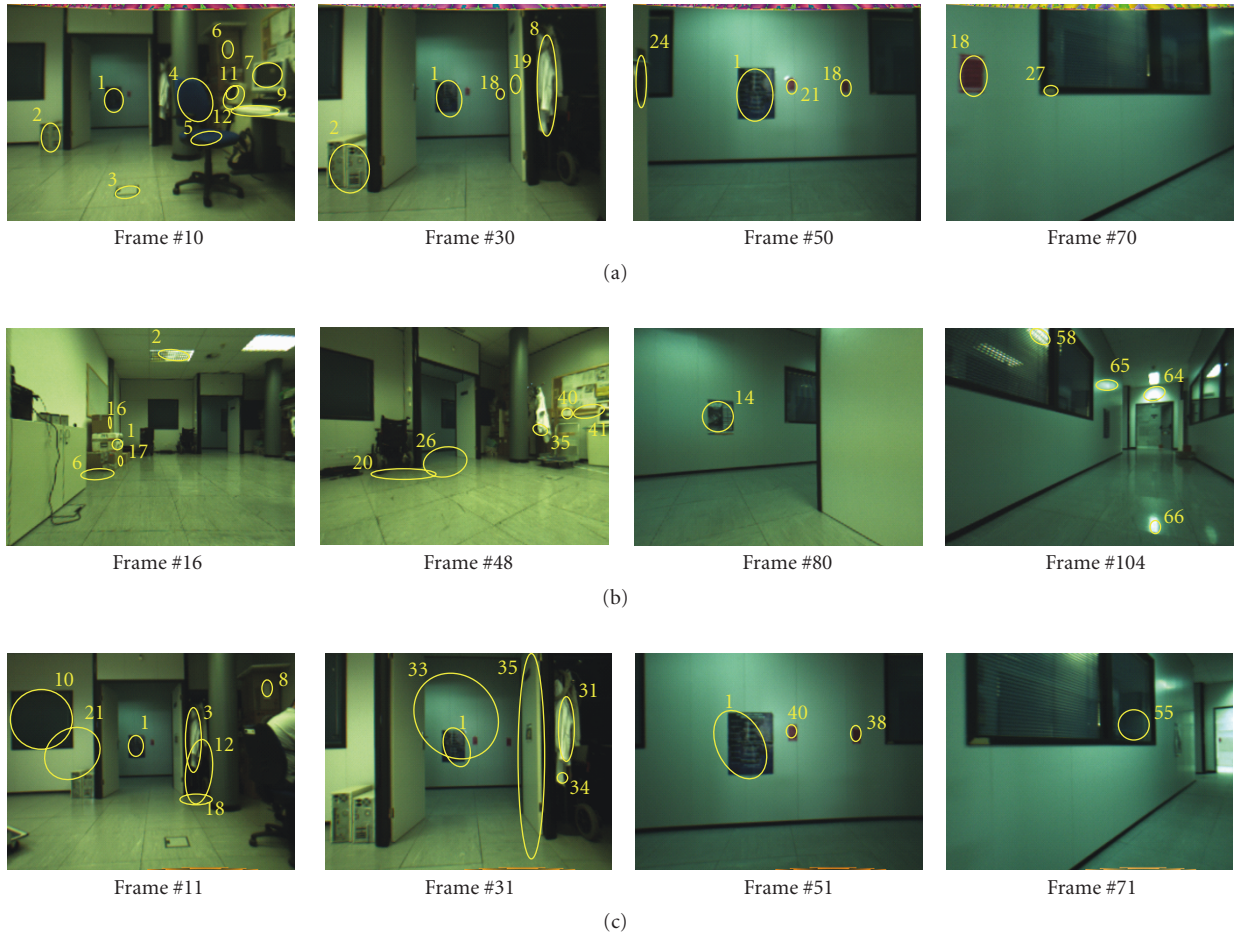


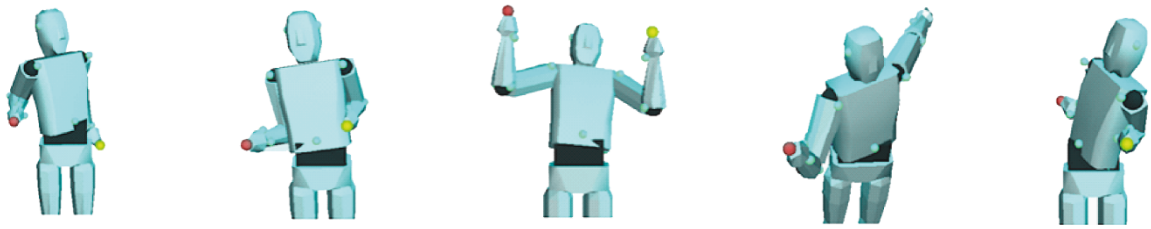
Figure 4: Visual landmarks detection results: (a) frames of video sequence #1, (b) frames of video sequence #2, and (c) frames of video sequence #3. Representing ellipses have been chosen to have the same first and second moments as the originally arbitrarily shaped region (matched landmarks inside of the same video sequence have been marked with the same index).

Table 1: Gestures used to test the system

Gesture	Description
Left up	Point up using the left hand
Left	Point left using the left hand
Right up	Point up using the right hand
Right	Point right using the right hand
Right forward	Point forward using the right hand
Stop	Move left and right hands forward
Hello	Wave the right hand
Hands up	Move left and right hands up

7.3. Testing the Approach in a Human Gesture Recognition Framework. The experiments performed to test the human gesture recognition stage involved different demonstrators executing different gestures in a noncontrolled environment. These users performed various executions of the upper-body social gestures listed in Table 1. No specific markers nor special clothes were used. As the stereo system has a limited range, the demonstrator was told to stay at a distance close to 1.7 meters from the cameras. The pre-attention stages provides this module with a set of skin-coloured regions. From this set of proto-objects, the faces of tentative human demonstrators are detected using a cascade detector based on the scheme proposed by Viola and Jones (see [30] for details). Once faces are detected, the closest face is chosen. The silhouette of the human demonstrator can be obtained by using a fast connected component algorithm that takes into account the information provided by the 3D position of the selected face. Human hands are detected as the two biggest skin colour regions inside this silhouette. It must be considered that this silhouette may also contain objects that are close to the human. The recognition system is then executed to identify the performed gesture. Figure 5 shows human heads and hands obtained when the gesture recognition system is executed on the previously described system. As depicted, the system is able to detect human faces in the field of view of the robot and it is also able to capture the upper-body motion of the closer human at human interaction rates.

(a)



(b)

FIGURE 5: Human motion capture results: (a) left image of the stereo pair with head (yellow) and hands (green) regions marked, and (b) 3D model showing captured pose.

8. Conclusions and Future Work

This paper has presented a visual attention model that integrates bottom-up and top-down processing. It runs at 15 frames per second using 320×240 images on a standard Pentium personal computer when there are less than five inhibited (tracked) proto-objects. The model accomplishes two selection stages, including a semiattentive computation stage where the inhibition of return has been performed and where a list of attended proto-objects is stored. This list can be used as a working memory, being employed by the behaviors to search for proto-objects which share some desired features. At the pre-attentive stage, the visual scene is divided into perceptually uniform blobs. Thus, the model can direct the attention on proto-objects, similarly to the behavior observed in humans. In order to deal with dynamic scenarios, the inhibition of return is performed by tracking the proto-objects. Specifically, this work uses the mean-shift tracker. Finally, this attention mechanism is integrated with an attentive stage that will control the field of attention following two different behaviors. The first behavior is a visual perception system which main goal is to help in the learning process of a social robot. The second one is a system to autonomously acquire visual landmarks for mobile robot simultaneous localization and mapping. We do not discuss in this paper the way these behaviors emerge or how the task-dependent parameters of the model are learnt. These issues will constitute our main future work.

Acknowledgments

This work has been partially granted by the Spanish MICINN and FEDER funds project no. TIN2008-06196 and by the Junta de Andalucía project no. P07-TIC-03106.

References

[1] J. Duncan, "Selective attention and the organization of visual information," *Journal of Experimental Psychology*, vol. 113, no. 4, pp. 501–517, 1984.

[2] B. J. Scholl, "Objects and attention: the state of the art," *Cognition*, vol. 80, no. 1-2, pp. 1–46, 2001.

[3] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, no. 1, pp. 77–123, 2003.

[4] F. Orabona, G. Metta, G. Sandini, and F. Sandoval, "A proto-object based visual attention model," in *Proceedings of the 4th International Workshop on Attention in Cognitive Systems (WAPCV '07)*, L. Paletta and E. Rome, Eds., vol. 4840 of *Lecture Notes in Computer Science*, pp. 198–215, Springer, Hyderabad, India, 2007.

[5] C. R. Olson, "Object-based vision and attention in primates," *Current Opinion in Neurobiology*, vol. 11, no. 2, pp. 171–179, 2001.

[6] Y. Yu, G. K. I. Mann, and R. G. Gosine, "An Object-Based Visual Attention Model for Robotic Applications," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 40, no. 3, pp. 1–15, 2010.

[7] S. Frintrop, G. Backer, and E. Rome, "Goal-directed search with a top-down modulated computational attention system," in *Proceedings of the 27th Annual Meeting of the German Association for Pattern Recognition (DAGM '05)*, W. G. Kropatsch, R. Sablatnig, and A. Hanbury, Eds., vol. 3663 of *Lecture Notes in Computer Science*, pp. 117–124, Springer, Vienna, Austria, 2005.

[8] A. Dankers, N. Barnes, and A. Zelinsky, "A reactive vision system: active-dynamic saliency," in *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS '07)*, 2007.

[9] G. Backer and B. Mertsching, "Two selection stages provide efficient object-based attentional control for dynamic vision," in *Proceedings of the International Workshop on Attention and Performance in Computer Vision (WAPCV '03)*, pp. 9–16, Springer, Graz, Austria, 2003.

[10] Z. W. Pylyshyn, "Visual indexes, preconceptual objects, and situated vision," *Cognition*, vol. 80, no. 1-2, pp. 127–158, 2001.

[11] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.

[12] M. Z. Aziz, *Behavior adaptive and real-time model of integrated bottom-up and top-down visual attention*, Ph.D. thesis, Fakultät für Elektrotechnik, Informatik und Mathematik, Universität Paderborn, 2000.

[13] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[15] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[16] P. Neri, "Attentional effects on sensory tuning for single-feature detection and double-feature conjunction," *Vision Research*, vol. 44, no. 26, pp. 3053–3064, 2004.

[17] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: an alternative to the feature integration model for visual search," *Journal of Experimental Psychology*, vol. 15, no. 3, pp. 419–433, 1989.

[18] J. M. Wolfe, "Guided Search 2.0: a revised model of visual search," *Psychonomic Bulletin and Review*, vol. 1, pp. 202–238, 1994.

[19] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, 2005.

[20] A. Borji, M. N. Ahmadabadi, B. N. Araabi, and M. Hamidi, "Online learning of task-driven object-based visual attention control," *Image and Vision Computing*, vol. 28, no. 7, pp. 1130–1145, 2010.

[21] R. Marfil, A. Bandera, A. Bandera, and F. Sandoval, "Comparison of perceptual grouping criteria within an integrated hierarchical framework," in *Proceedings of the Graph-Based Representations in Pattern Recognition (GbRPR '09)*, A. Torsello and F. Escolano, Eds., vol. 5534 of *Lecture Notes in Computer Science*, pp. 366–375, Springer, Venice, Italy, 2009.

[22] R. Marfil, L. Molina-Tanco, A. Bandera, J. A. Rodríguez, and F. Sandoval, "Pyramid segmentation algorithms revisited," *Pattern Recognition*, vol. 39, no. 8, pp. 1430–1451, 2006.

[23] J. Huart and P. Bertolino, "Similarity-based and perception-based image segmentation," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '05)*, pp. 1148–1151, September 2005.

[24] R. Vázquez-Martín, R. Marfil, P. Núñez, A. Bandera, and F. Sandoval, "A novel approach for salient image regions detection and description," *Pattern Recognition Letters*, vol. 30, no. 16, pp. 1464–1476, 2009.

[25] R. Marfil, L. Molina-Tanco, A. Bandera, and F. Sandoval, "The construction of bounded irregular pyramids using a union-find decimation process," in *Proceedings of the Graph-Based Representations in Pattern Recognition (GbRPR '07)*, F. Escolano and M. Vento, Eds., vol. 4538 of *Lecture Notes in Computer Science*, pp. 307–318, Springer, Alicante, Spain, 2007.

[26] L. Itti, "Real-time high-performance attention focusing in outdoors color video streams," in *Human Vision and Electronic Imaging (HVEI '02)*, vol. 4662 of *Proceedings of SPIE*, pp. 235–243, 2002.

[27] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[28] G. Backer, B. Mertsching, and M. Bollmann, "Data- and model-driven gaze control for an active-vision system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1415–1429, 2001.

[29] J. P. Bandera, A. Bandera, L. Molina-Tanco, and J. A. Rodríguez, "Vision-based gesture recognition interface for a social robot," in *Proceedings of the Workshop on Multimodal Human-Robot Interfaces (ICRA '10)*, 2010.

[30] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[31] K. Mikolajczyk, T. Tuytelaars, C. Schmid et al., "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.