Applied Bionics and Biomechanics 9 (2012) 409–425
DOI 10.3233/ABB-2011-0030
IOS Press

Visual attention mechanism for a social robot

Juan Pedro Bandera[a,*], R. Marfil[a], Antonio Jesús Palomino[a], Ricardo Vázquez-Martín[b] and Antonio Bandera[a]
[a]*ISIS Group, Dpto. Tecnología Electrónica, University of Málaga, Málaga, Spain*
[b]*Centro Andaluz de Innovación y Tecnologías de la Información y las Comunicaciones, Málaga, Spain*

Abstract. This paper describes a visual perception system for a social robot. The central part of this system is an artificial attention mechanism that discriminates the most relevant information from all the visual information perceived by the robot. It is composed by three stages. At the preattentive stage, the concept of saliency is implemented based on 'proto-objects' [37]. From these objects, different saliency maps are generated. Then, the semiattentive stage identifies and tracks significant items according to the tasks to accomplish. This tracking process allows to implement the 'inhibition of return'. Finally, the attentive stage fixes the field of attention to the most relevant object depending on the behaviours to carry out. Three behaviours have been implemented and tested which allow the robot to detect visual landmarks in an initially unknown environment, and to recognize and capture the upper-body motion of people interested in interact with it.

Keywords: Social robots, active vision, attention mechanism, human-robot interaction, visual landmark detection

1. Introduction

People exhibit a robust ability to extract the relevant information from perceived scenarios. This ability allows people to execute many different, complex behaviours such as to navigate in a huge variety of initially unknown environments or to interact with other people and appropriately interpret their behaviours. Developing computational perception systems that emulate this ability becomes a critical step in designing 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 navigate in initially unknown environments or to grasp an object. In order to accomplish these tasks it is typically assumed that the perception system of a robot should imitate the ability of natural vision systems to select the most salient information from the broad visual input. However, this selective system is commonly developed as a task-independent process. Therefore, the focus of attention must be sequentially moved to every detected relevant region in order to extract from the scene the information that the robot needs to satisfy a specific task. The complexity of this gathering process makes difficult for the robot to solve a certain task while it is interacting with people. This issue should be addressed as some of the new applications for robots require them to cooperate with people as socially interactive partners [19], needing a fast response of the robot to a huge variety of different stimulus.

This proposal describes a visual perception system for a social robot. A social robot is an autonomous agent which is not only able to navigate and solve other common tasks, but also it is able to communicate and interact with people and other social robots. This implies that the social robot must simultaneously perceive a great variety of natural social cues from

*Corresponding author: Juan Pedro Bandera, ISIS Group, Dpto. Tecnología Electrónica, University of Málaga, Málaga, Spain. E-mail: jpbandera@uma.es.

visual and auditory channels, and must reply to these stimulus at human rates. Specifically, the proposed perception system will be used to extract, from the visual input data, the information that the robot will need to accomplish both navigation and human-robot interaction behaviours. Besides it is interesting, to achieve an intuitive interaction with people, that the robot is able to perceive the real world in a similar way that people do. Thus, the socially interactive robot will interpret the same phenomena that people observe [17]. To accomplish both requirements, the central element of our proposal is an object-based visual attention mechanism which will be able to discriminate, from all the low-level information provided by the robot's cameras, the most relevant data useful to fulfill the currently executed tasks.

1.1. Related work

In biological vision systems, the attention mechanism is the responsible of selecting the relevant information from the sensed field of view so that the complete scene can be analyzed using a sequence of rapid eye saccades [2]. This attention behaviour has been imitated by artificial vision systems in order to optimize computational resources. Probably one of the most influential theoretical models of visual attention is the spotlight metaphor [18], that has inspired many concrete computational models [24, 23, 16]. These approaches are related with the *feature integration theory*, a biologically plausible theory proposed to explain human visual search strategies [45]. According to this model, these attention mechanisms are organized into two main stages. First, in a preattentive task-independent stage, a number of parallel channels compute image features. The extracted features are integrated into a single saliency map which codes the saliency of each image region. The most salient regions are selected from this map. Second, in an attentive task-dependent stage, the spotlight is moved to each salient region to analyze it in a sequential process. Analyzed regions are included in an *inhibition map* to avoid the spotlight moving to an already visited region. Thus, while the second stage must be redefined for different systems, the preattentive stage is general for any application. Although these models have good performance in static environments, they cannot in principle handle dynamic environments due to their impossibility to take into account the motion and the occlusions of the objects in the scene. In order to solve this problem, [27] propose an attention mechanism which incorporates depth and motion as features for the computation of saliency.

The previously described methods deploy attention at the level of space locations (*space-based models of visual attention*). 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. Only if different visual features, which constitute the same object, come from the same region of space, an attention shift will not be required [42]. On the contrary, other approaches deploy attention at the level of objects instead to a generic region of space. *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. These models reflect the fact that the perception abilities must be optimized to interact with objects and not just with disembodied spatial locations. Thus, visual systems that follow this approach will segment complex scenes into objects which can be subsequently used for recognition and action. However, recent phsycological research shows that, in natural vision, the preattentive process divides a visual input into raw or primitive objects [37] instead of well-defined objects. Some authors use the notion of *proto-objects* [38] to refer to these primitive objects, that are defined as units of visual information that can be bound into a coherent and stable object.

Nevertheless, space-based and object-based approaches are not mutually exclusive, and several researchers have proposed attentional models that integrate both approaches. Thus, [42] combine object-based and feature-based theories in the model of visual attention. In its current form, this model is able to replicate human viewing behaviour. However, it needs input images to be manually segmented. That is, it uses information that is not available in a preattentive stage, before objects are recognized [38]. Another approach following the space- and object-based integration is the one proposed by [11], which employs a Bayesian model to describe the visual attention mechanism.

In the last years, efforts in artificial attention field are oriented to enhance the attention capabilities of the different models. For example, [3] describe a framework to integrate visual attention with a spatial memory derived from a 3D simulator. This model allows to connect the perception system with a virtual reality environment. Other approaches choose to expand the visual field of their systems. In this way, a spherical visual attention model based on a omnidirectional sensor is proposed in [12]. Finally, some models build their saliency maps taking into account not only visual information but also using another senses (e.g. [40] integrate visual and acoustic cues to improve the exploration behaviour of a humanoid robot).

1.2. Overview of the proposed perception system

This paper presents an object-based model of visual attention for a social robot which moves in a dynamic scenario. Its main contribution is the integration into the same system of different components related to visual attention. These components allow the system to achieve bottom-up (data-driven) and top-down (model-driven) processing. The bottom-up component determines and selects salient 'proto-objects' by integrating different features into the same hierarchical structure. These 'proto-objects' [38, 39] are image entities which do not necessarily correspond with a recognizable object, although they posses some of the characteristics of objects. Thus, it can be considered that they are the result of the initial segmentation of the image input into candidate objects or *segmented perceptual units* [38]. On the other hand, the top-down component moves the focus of attention to a certain object depending on its saliency value and on the task to acomplish. Finally, in a dynamic scenario, 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 tracking approach for 'inhibition of return' is employed [32]. The 'proto-objects' selection and the tracking process will be conducted using the same hierarchical structure: the Bounded Irregular Pyramid (BIP) [28, 31].

Figure 1 shows an overview of the proposed attention mechanism. Following our previous works [32], it is composed by three main modules. The first one implements a concept of saliency based on 'proto-objects'. The selection of these 'proto-objects' is conducted from an image segmentation process which results are closely related to the ones obtained by humans, as studies using the Berkeley Segmentation Dataset and Benchmark (BSDB) have demonstrated [33, 30]. From these objects, different saliency maps are generated. These maps are the input of the semiattentive stage. This semiattentive stage deals with dynamic scenarios and considers the current task in the selection of the 'proto-objects'. Thus, the selection of a set of sensed data as focus of attention will not only depend on its saliency value but also on the tasks to reach. This semiattentive stage also implements the 'inhibition of return', i.e. the process which avoids attention being immediately directed to a previously attended 'proto-object'. This local inhibition is achieved by tracking the set of relevant 'proto-objects'. The attentive stage fixes the field of attention to the most salient 'proto-object' depending on the current behaviour.

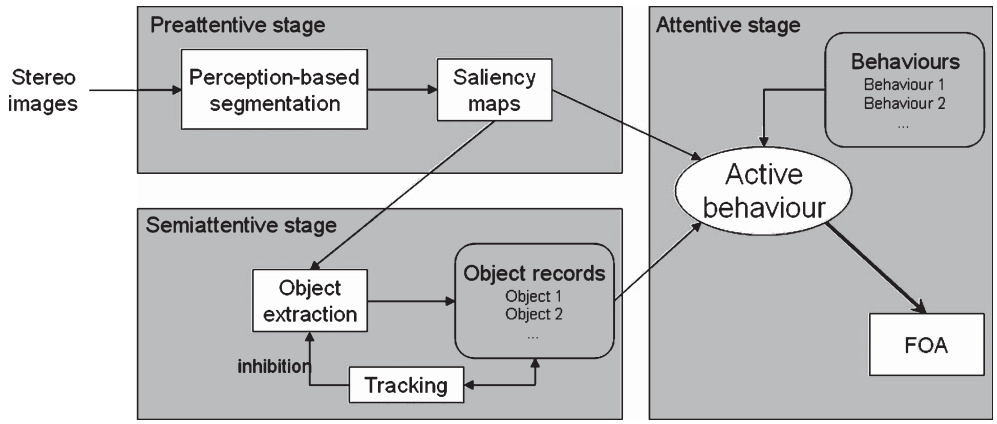

Fig. 1. Overview of the proposed attention mechanism.

In order to provide to a social robot with the necessary abilities to autonomously navigate and interact with people in a dynamic scenario, three behaviours have been included in the proposed attention mechanism: a human face detection and recognition behaviour, a human gesture recognition behaviour and a visual natural landmark detection behaviour. These behaviours are the responsible to recognize a person who is interested in establishing an interaction, and to provide visual natural landmarks for mobile robot navigation. Switching between these behaviours lays in higher level decision layers that are beyond the scope of this paper.

The remainder of the paper is organized as follows: Section 2 presents a description of the computation of the 'proto-objects' and their associated saliency values. The scheme used to implement the inhibition of return is described in Section 3. Section 4 presents the attentive stage. In this paper, three different behaviours have been implemented. These behaviours are focused on recognizing a detected person and capturing its upper-body motion, and on extracting distinguished visual natural landmarks. Section 5 deals with some relevant experimental results. Finally, conclusions and future works are presented in Section 6.

2. Preattentive stage: Saliency maps computation

The proposed visual attention model employs a concept of salience based on 'proto-objects'. These objects are defined as the union of a set of blobs of uniform color and disparity of the image which will be partially or totally bounded by the edges obtained using a Canny detector. The proposed method used to obtain these entities has two main stages. In the first stage the input image pixels are grouped into blobs of uniform color in a pre-segmentation process. These blobs preserve the image geometric structure as each significant feature contains at least one blob. In the second stage or perceptual grouping stage, these blobs are grouped into a smaller set of 'proto-objects', taking into account not only the internal visual features of the blobs but also the external relationships among them.

A 'preattentive object' catches the attention if it differs from its immediate surrounding. In contrast with other previous works [5, 38] which only compute one saliency map in the preattentive stage, the proposed approach computes two different saliency maps associated to the set of 'proto-objects' previously extracted. The idea is to be able to select different salient 'proto-objects' depending on the current behaviour or task. Hence, there are two different saliency maps. In the first one, the color and luminosity contrasts between the 'preattentive object' and all the regions in its surrounding and its distance to the robot are evaluated. The most relevant 'proto-objects' in this map are the most contrasted and closest ones. The second one evaluates if the preattentive object is skin colored. The aim is that the closest skin colored ones will be the most salient 'proto-objects'. This second saliency map will allow a social robot to select from the scene the regions where a human face, and thus a potential interaction partner, can be probably located.

2.1. Implementation

2.1.1. 'Preattentive object' detection

As it has been aforementioned, the proposed approach firstly segments the image into perceptually uniform blobs, and then it groups these blobs taking into account a more complex criterion. The final set of regions constitute the set of 'proto-objects'. Both stages are performed using the Bounded Irregular Pyramid (BIP), an irregular pyramid which has demonstrated to be able to perform fast segmentation of color images [31].

The Bounded Irregular Pyramid (BIP) [28, 31] is a mixture of regular and irregular pyramids: a $2 \times 2/4$ regular structure is used in the homogeneous regions of the input image and a simple graph structure in the non-homogeneous ones. The mixture of both structures generates an irregular configuration which is described as a graph hierarchy in which each level $G_l = (N_l, E_l)$ consists of a set of nodes, N_l, linked by a set of intra-level edges E_l. Each graph G_{l+1} is built from G_l by computing the nodes of N_{l+1} from G_l and establishing the inter-level edges $E_{l,l+1}$. Therefore, each node n_i of G_{l+1} has associated a set of nodes of G_l, which is called the *reduction window* of n_i. This includes all nodes linked to n_i by an inter-level edge. The node n_i is called *parent* of the nodes in its reduction window, which are called *children*. The successive levels of the hierarchy are built using a regular decimation process and an union-find strategy [31]. Therefore, there are two types of nodes: nodes belonging to the $2 \times 2/4$ structure, named regular nodes, and virtual nodes or nodes belonging to the irregular structure. In any case, two nodes $n_i \in N_l$ and $n_j \in N_l$ which are neighbors at level l are linked by an intra-level edge $e_{ij} \in E_l$.

The proposed approach uses a BIP structure to accomplish the detection of the 'proto-objects' and the subsequent computation of the saliency maps. In this hierarchy, the first levels perform the pre-segmentation stage using a distance based on color to group pixels into perceptually homogeneous blobs. In order to introduce low-level information within the BIP, all the nodes of the structure have three parameters associated: saturation S, hue H and value V of the HSV color space. All the parameters of a node n at level l are equal to the average of parameters of the nodes in its reduction window, i.e. the nodes of the level l-1 which are linked to n. The BIP structure is built based on the concept of similarity between nodes. Two nodes n_i and n_j are similar if the distance between their HSV values is less than a threshold T.

The graph $G_0 = (N_0, E_0)$ is a 8-connected graph where the nodes are the pixels of the original image. The parameters of the nodes in $G_0 = (N_0, E_0)$ are equal to the parameters of their corresponding image pixels. The process to build the graph $G_{l+1} = (N_{l+1}, E_{l+1})$ from $G_l = (N_l, E_l)$ is briefly described below (see [31] for further details):

1) *Regular decimation process*. If four regular neighbor nodes of the level l have similar color, they are grouped together, generating a regular node in $l + 1$.
2) *Parent search and intra-level twining*. Once the regular structure is generated, there are some regular orphan nodes (regular nodes without parent). From each of these nodes (i, j, l), a search is made for the most similar node with parent in its neighborhood $\xi_{(i,j,l)}$. If this neighbor node is found, the node (i, j, l) is linked to the parent of this neighbor node. On the contrary, if for this node a parent is not found, then a search is made for the most similar neighbor node without parent to link to it. If this node is found, then both nodes are linked, generating a virtual node at level $l + 1$.
3) *Virtual parent search and virtual node linking*. Each virtual orphan node n_i searches for the most similar node with parent in its neighborhood ξ_{n_i}. If for n_i a parent is found, then it is linked to it. On the other hand, if a parent is not found, the virtual orphan node n_i looks for the most similar orphan node in its neighborhood to generate a new virtual node at level $l + 1$. The only restriction to this step is that a virtual node cannot be linked to a regular parent. This procedure allows to preserve the regular nature of the regular part of the BIP.

The hierarchy stops growing when it is no longer possible to link together any more nodes because they are not similar. In order to perform the pre-segmentation, the orphan nodes are used as roots. The described method has been tested and compared with other similar pyramid approaches for color image segmentation [28]. This comparative study concludes that the BIP runs faster than other irregular approaches when benchmarking is performed in a standard sequential computer. Besides, the BIP obtains similar results than the main irregular structures. Figure 2b shows the pre-segmentation images associated to the images in Fig. 2a. It can be noted as the input images are correctly segmented into blobs of uniform colour.

After the local similarity pre-segmentation stage, grouping blobs aims at simplifying the content of the obtained partition in order to obtain the set of final 'proto-objects'. This process is also performed using the BIP structure: the roots of the pre-segmented blobs are considered as virtual nodes which constitute the first level of the perceptual grouping multi-resolution output. Successive levels can be built using the virtual parent search and virtual node linking scheme previously described. This perceptual grouping process

Fig. 2. a) Original image; b) pre-segmented blobs; c) obtained regions after the perceptual grouping; and d) color contrast saliency map.

is explained in detail in [30]. The main aspect is the use of a similarity distance that complements a color contrast measure with internal regions properties and with attributes of the boundary shared by both regions. Therefore, this distance has three main components: the color contrast between image blobs, the edges of the original image computed using the Canny detector and the disparity of the image blobs obtained from stereo information. Then, the distance between two nodes n_i and n_j is defined as

$$\Upsilon(n_i, n_j) = \sqrt{w_1 \cdot \left(\frac{d(n_i, n_j) \cdot b_i}{\alpha \cdot (c_{ij}) + (\beta \cdot (b_{ij} - c_{ij}))} \right)^2 + w_2 \cdot (disp(n_i) - disp(n_j))^2} \quad (1)$$

where $d(n_i, n_j)$ is the color distance between n_i and n_j and $disp(x)$ is the mean disparity associated to the base image region represented by node x. b_i is the perimeter of n_i, b_{ij} is the number of pixels in the common boundary between n_i and n_j and c_{ij} is the set of pixels in this common boundary which corresponds to pixels of the boundary detected by the Canny detector. α and β are two constant values used to control the influence of the Canny edges in the grouping process. We set these parameters to 0.1 and 1.0, respectively. In the same way w_1 and w_2 are two constant values which weight the terms associated with the color and the disparity. In our case they are set to 0.5 and 1.0, respectively.

A threshold value T_{perc} is used to discriminate between similar and not similar blobs. The grouping process is iterated until the number of nodes remains constant between two successive levels.

After the pre-segmentation and perceptual grouping stages, the nodes of the BIP with no parent will be the roots of the 'preattentive objects'. It must be appreciated that these 'proto-objects' can be represented as hierarchical structures, where the object root constitutes the higher level of the representation and the nodes of the input image linked to this root conform its lower level.

Figure 2c shows the set of 'proto-objects' associated to the images in Fig. 2a. It can be noted that, although the number of obtained regions have not been significantly reduced with respect to the number of blobs of the preattentive stage, they provide an image segmentation which is more coherent with the human-based image decomposition. This fact has been proven by evaluating the proposed perceptual grouping approach using the Berkeley Segmentation Dataset and Benchmark [33]. Obtained results improve the ones obtained in our previous work [30], and equal the better proposals to date [26], due to the inclusion of the Canny edges in the process. It must be noted that, because of the used images, the disparity value has not been included in this evaluation.

2.1.2. Saliency maps computation

In order to compute the skin color saliency map, two features of the 'proto-objects' are taken into account: its possibility to be skin colored and its disparity. In order to determine if a 'preattentive object' is skin colored, the skin chrominance model proposed by [43] has been used. Once the chrominance model has been established, the steps to detect the skin colored 'preattentive objects' are the following: first, the perceptually segmented RGB image is transformed into a TSL image. Second, the Mahalanobis distance from the color of each 'preattentive object' to the mean vector of the skin chrominance model is computed. If this distance is less than a threshold T_s then the 'preattentive object' is marked with a value of 255 in the skin color feature map (SKF). In any other case, it is set to 0. The disparity D_i of each 'proto-object' R_i is obtained by averaging the disparity values associated with each pixel of the 'proto-object'.

The final skin color saliency value (SKS_i) of each 'proto-object' is obtained as a single summation of both skin color and disparity features:

$$SKS_i = \frac{SKF_i + D_i}{2} \quad (2)$$

An example of this map is shown in Fig. 3.

We compute the color contrast saliency value of each 'proto-object' using color contrast and intensity con-

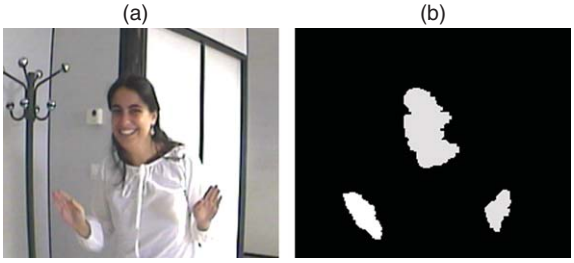

Fig. 3. a) Original image; b) skin color saliency map SKS.

trast measures as well as the disparity associated to it. As each of these 'proto-objects' corresponds to a root of the BIP structure previously generated, it contains all the necessary information about the concerned image region, such as its average chromatic phasor and intensity and the set of its neighbors.

Then, we compute the color contrast of a 'preattentive object' $\mathcal{R}_i$ as the mean color gradient MCG_i along its boundary to the neighbor blobs:

$$MCG_i = \frac{S_i}{b_i} \sum_{j \in N_i} b_{ij} * d(\langle C_i \rangle, \langle C_j \rangle) \tag{3}$$

being b_i the perimeter of $\mathcal{R}_i$, N_i the set of regions which are neighbors of $\mathcal{R}_i$, b_{ij} the length of the perimeter of the region $\mathcal{R}_i$ in contact with the region $\mathcal{R}_j$, $d(\langle C_i \rangle, \langle C_j \rangle)$ the color distance between the color mean values $\langle C \rangle$ of the regions $\mathcal{R}_i$ and $\mathcal{R}_j$ and S_i the mean saturation value of the region $\mathcal{R}_i$.

The use of S_i in the MCG avoids that color regions with low saturation (grey regions) obtain a higher value of color contrast than pure color regions. The problem is that white, black and pure grey regions are totally suppressed. To take into account these regions, the luminosity contrast is computed. The luminosity contrast of a region $\mathcal{R}_i$ is the mean luminosity gradient MIG_i along its boundary to the neighbor regions:

$$MLG_i = \frac{1}{b_i} \sum_{j \in N_i} b_{ij} * d(\langle I_i \rangle, \langle I_j \rangle) \tag{4}$$

being $\langle I_i \rangle$ the mean luminosity value of the region $\mathcal{R}_i$.

Then the final color contrast value of $\mathcal{R}_i$ is computed as:

$$MG_i = \sqrt{MCG_i^2 + MLG_i^2} \tag{5}$$

Finally the color contrast saliency map which combines color contrast and disparity information is computed as follows:

$$MGS_i = \frac{MG_i + D_i}{2} \tag{6}$$

Figure 2d shows the color contrast saliency map MGS_i obtained from the image in Fig. 2a.

3. Semiattentive stage: Inhibition of return

Human visual psychophysics studies have demonstrated that a local inhibition is activated in the saliency map to avoid attention being immediately directed to a previously attended region. In the context of artificial models of visual attention, this 'inhibition of return' has been usually implemented using a 2D inhibition map, that contains suppression factors for one or more focuses of attention that were recently attended [22, 20]. However, this 2D inhibition map is not able to handle motion of inhibited objects nor motion of the vision system itself. Dynamic scenes require totally different processes to be handled in comparison to static scenes, because the locations and shapes of the objects may change due to motion and minor illumination differences between consecutive frames. In this situation, establishing a correspondence from regions of the previous frame to those of the next frame becomes a significant issue.

In order to allow tracking an object while it changes its location, the model proposed by [4] relates the inhibitions to features of activity clusters. However, the scope of dynamic inhibition becomes very limited as it is related to activity clusters rather than objects themselves [2]. Thus, it is a better option to attach the inhibition to moving objects [44]. For instance, the recent proposal of [2] utilizes a queue of inhibited region features to maintain inhibition in dynamic scenes.

The proposed system implements an object-based 'inhibition of return' [32]. A list of the last attended 'preattentive objects' is maintained at the semiattentive stage of the visual attention model. This list stores information about the color and last position of the 'preattentive object'. It also stores the last hierarchical representation associated to each 'preattentive object'. When the vision system moves, the proposed approach keeps track of the 'proto-objects' that it has already visited.

3.1. Implementation

The employed tracking algorithm is also based on the Bounded Irregular Pyramid (BIP)and it is detailed explained in [29]. This tracking algorithm allows a fast tracking of non-rigid objects, while it does not need a previous learning of different object views. This is possible due to the use of weighted templates which follow up the viewpoint and appearance changes of the objects to track. The templates and the targets are represented using BIPs. Thus, the generation of the whole set of 'proto-objects' and the tracking of the attended ones to inhibit them are performed into

the same hierarchical framework. This will allow to reduce the computation time associated to the whole model of visual attention.

4. Attentive stage

The attentive stage is the responsible of executing the different behaviours that the social robot exhibits. The system presented in this paper implements three example behaviours for the attentive stage, that allow a social robot to autonomously navigate and interact with people in a dynamic scenario. These behaviours deal with face recognition, human motion capture and environmental landmarks detection.

It is important to consider that the process of selecting the most adequate behaviour for each situation lays in higher decision levels, that are not addressed in this paper. Instead, it presents a framework where different behaviours are executed in order to depict the potential of the proposed architecture. In any case, for the three behaviours used in this paper in the attentive stage, a tentative simple switching mechanism is proposed. This mechanism executes the natural landmark detection behaviour until a skin colour 'proto-object' with a high value of saliency is detected. This event makes the robot stop identifying visual landmarks and focus on searching for a human face. If no faces are present, then the visual landmark detection behaviour is launched again. However, if a face is found, the face recognition and the human motion capture behaviours are executed consecutively in order to identify the human and his/her motion.

4.1. Face recognition

A social robot that works in real environments should be able to discriminate between the people in its surroundings, so that interaction can benefit from previous knowledge about these potential partners [13]. The first step towards allowing a robot to establish individualized communication channels is to provide it with the ability to recognize people. In the proposed system system, the face recognition behaviour implements this ability. In this behaviour only the skin colour saliency map is taken into account. It firstly visits all skin color 'proto-objects' from the skin colour saliency map in order to locate a human face. Once the faces have been located in the image they are recognized.

4.1.1. Implementation

In the proposed system, a method based on the use of Principal Component Analysis (PCA) and eigen objects is used for face recognition. This extracts low-dimensional subspaces which helps to simplify tasks such as classification [9]. The Karhunen Loeve Transform (KLT) and PCA are the eigenvector based techniques used for dimensionality reduction and feature extraction in automatic face recognition [9]. These techniques are used to create an eigenspace for the set of people to be recognized, in a previous training phase.

Once the eigenspace has been computed, each detected face can be represented by a set of decomposition coefficients obtained using the eigen objects associated to each person in the database. Then, these coefficients are used to compute the projections of the input face over each stored face. Projections are compared with the original input face, and the best match is set as the recognized person if their similarity is over a certain threshold. More details about the particular implementation of this face recognition process can be found in [9].

4.2. Human motion capture

In the current implementation of a simple switching among behaviours, this behaviour is executed only if the face recognition behaviour has been previously performed. Thus, although the Human Motion Capture (HMC) behaviour is not affected by the results of the face recognition behaviour (the human motion is captured regardless the human has been recognized or not), it assumes at least one human face is in the field of view of the cameras. The flow diagram of the HMC behaviour is shown in Fig. 4. Among the set of recognized faces by the face recognition behaviour, the most salient is selected and its 3D position is used to extract the upper-body pose of the perceived human. There are different options to perform this task from stereo images [36]. The proposed system uses a model-based approach, that requires to extract the silhouette from disparity maps, and to locate the 3D position of the hands. Once these data are obtained, silhouette information is used to estimate torso rotation and bending angles. The 3D position of the hands with respect to the face, on the other hand, is used to pose human arms. The use of this model-based method avoids local minima, a common problem of optimization and probabilistic methods [36]. Once an estimation of the human pose is obtained, this pose is translated to the

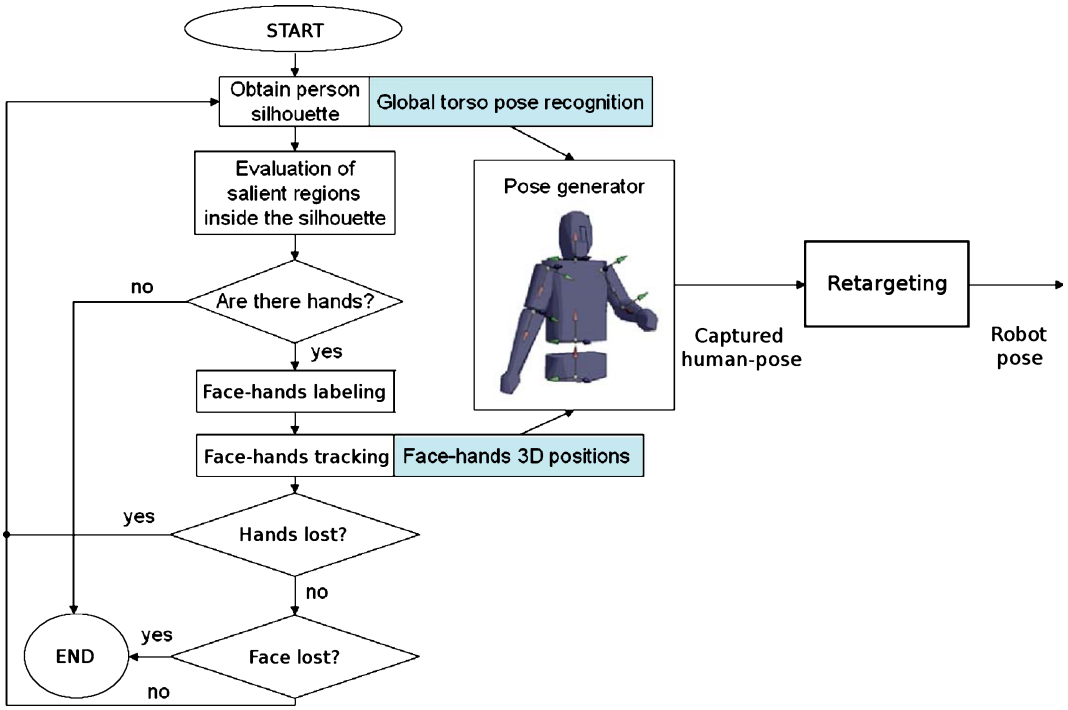

Fig. 4. Flow diagram of the HMC behaviour.

robot motion space using a retargeting module. The output of this module is the robot pose that imitates perceived human pose.

The HMC behaviour deals with partial occlusions and noisy data. It finishes when the face object is definitely lost.

4.2.1. Implementation

The HMC algorithm proposed in this paper is the integration of various contributions that have been already presented in previous work. Thus, the torso pose estimator is detailed in [15], and the method to pose arms, based on inverse kinematics, is described in [6]. As this HMC system is proposed to be used in a social robot, additional computation is required to translate, or retarget, the captured human motion to the robot motion space. This process is deeply described in [8]. The rest of this Section provides a brief description of all these algorithms, focusing on the elements that have been added or modified to ease integration.

Using the closest human face obtained from the set of faces recognized by the face recognition behaviour, the silhouette of this human is extracted using the following procedure [25]: the disparity map is thresholded taking into account that the maximum distance between the detected head and one hand of the same person is determined by the length of a stretched arm. We consider this length $L(arm)$ not to be greater than one meter. Then, depth values $d(i, j)$ that does not meet Eq. (7) are discarded:

$$U_d - L(arm) < d(i, j) < U_d + L(arm) \qquad (7)$$

being U_d the mean depth value of the face. Once this filtered disparity map is obtained, the silhouette of the person can be extracted from it using connected components. This silhouette is provided to the pose generator (Fig. 4) to extract torso bending and rotation angles. Then, following previous proposals [13], the two largest skin colour, non-face objects tracked inside the silhouette are labeled as the human hands, as depicted in Fig. 5d. In order to distinguish left and right hands, it is considered in the initial frames of the HMC behaviour that the human stands still in front of the robot, with the left hand located in the left part of the body, and the right hand located in the right part of the body.

The pose generator depicted in Fig. 4 uses a kinematic human model to translate the 3D positions

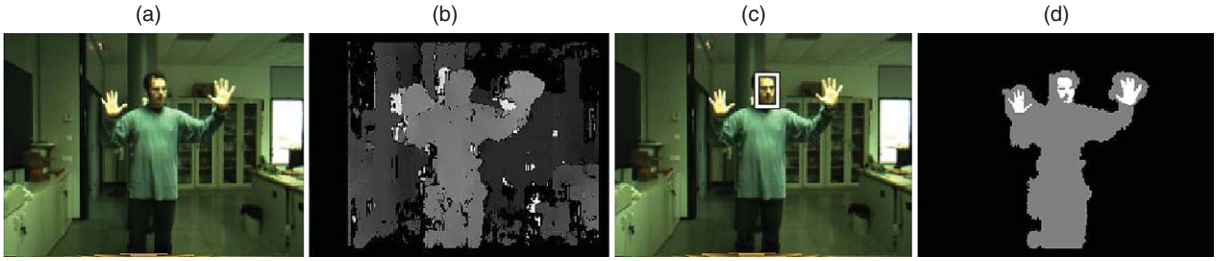

Fig. 5. HMC attentive stage: a) Left image of the input stereo pair; b) disparity map; c) detected faces; and d) obtained silhouette showing skin color objects associated to human head and hands.

of detected human items to a global correct pose. This translation is based on a fast analytic inverse kinematics algorithm running over a model that avoids incorrect poses. As we have restricted ourselves to capture upper-body motion, the geometric model contains parts that represent hips, head, torso, arms and forearms of the human to be tracked [6], Model proportions have been set to average human values, and model dimensions are scaled to fit the performer's height. The following steps are applied in order to obtain the final set of joint angles for frame k, $\overrightarrow{\theta(k)} = (\overrightarrow{\theta^t(k)}, \overrightarrow{\theta^l(k)}, \overrightarrow{\theta^r(k)}) = (\theta_0(k), \theta_1(k), \ldots, \theta_{10}(k))$, where $\overrightarrow{\theta^t(k)}$ corresponds to the three joint angles located in the torso and $\overrightarrow{\theta^l(k)}$, $\overrightarrow{\theta^r(k)}$ are the sets of joint angles located in the left an right arms, respectively:

- The first step is to locate the image region in which the torso of the human performer is most probably located. This process relies on the use of anthropometric tables [14] to estimate this region from human height, face position and waist position. In the proposed system the human begins interaction standing still in front of the robot. Thus, human height can be computed as $H = (1/0.92) \cdot h_{face}$, where h_{face} is the height of the face detected in the first frames [14]. Other relations extracted from anthropometric tables and used in this paper are detailed in [15]. As the person is not supposed to walk around while gesturing the robot, the waist position is set in these first frames. Waist position is then considered a fixed value during the rest of the interaction process. The face position is known for each frame.

  The anthropometric tables are used to provide a search region for the shoulders, the neck and the torso axis. The used anthropometric tables are valid for most adult people, being male or female [14]. However, the system would need different tables to perceive, for example, small children movements. Another option to adapt anthropometric values to different people is to execute a certain initialization phase in which standard values are adapted to the particular performer [1]. However, unsupervised initialization may be very difficult to achieve in real uncontrolled environments, in which noisy data and occlusions are present. On the other hand, manually performed initialization is not adequate for a system to be integrated in autonomous social robots.
- Once torso region has been delimited, the points that conform the medium axis are estimated using the following procedure: (i) For each row in the torso search region, the silhouette pixels are grouped into connected segments; (ii) the longest segment in the row is selected as the torso segment; (iii) the medium point in the torso segment is marked as a point of the medium axis; and (iv) once all the previous medium points have been extracted, the central limit theorem is applied to model the distribution of these points as a gaussian and filter outliers [15].

  This procedure reduces the influence of arms or other non-torso objects appearing in the torso search region. Once these points have been extracted, the projection of the torso medium axis is computed as the result of performing a 2D linear interpolation over all of them. Then, the depth information associated to the points in this line is used to compute the 3D position of torso medium axis.

  The medium axis allows to bend the torso. Torso rotations, on the other hand, are achieved by computing the depth difference between the two

shoulders for each frame. The shoulders 3D positions are estimated using anthropometric data and geometric relations [15]. Then, disparity values in the vicinity of these points are averaged to obtain the disparity of each shoulder. These estimated shoulder disparities are finally used to infer a 3D position for each shoulder, and the depth differences between these positions are used to rotate the torso [15].

Once torso bending and rotation angles have been computed, they are identified as the three components of $\overrightarrow{\theta^t(k)}$. The torso of the used model adopts these angles before the next steps of the HMC behaviour, that pose the arms from the 3D positions of the hand objects, are executed.

- In order to reduce the effects of disparity noise and outliers in the estimation of hand positions, a Gaussian filter is applied to the 3D positions of tracked left and right hands, $\overrightarrow{P^l(k)} = (P^l_x(k),\ P^l_y(k),\ P^l_z(k))$ and $\overrightarrow{P^r(k)} = (P^r_x(k),\ P^r_y(k),\ P^r_z(k))$.
- A constrained inverse kinematics (IK) algorithm is used to compute a set of joint angles for each arm of the kinematic model. This method obtains an arm pose that will put the hand of the model in the required position, as Eq. (8) depicts.

$$\begin{array}{ll} \overrightarrow{Q^l(k)} = f(\overrightarrow{P^l(k)}) & \overrightarrow{Q^r(k)} = f(\overrightarrow{P^r(k)}) \\ \overrightarrow{\theta^l(k)} = IK(\overrightarrow{Q^l(k)}) & \overrightarrow{\theta^r(k)} = IK(\overrightarrow{Q^r(k)}) \end{array} \tag{8}$$

where $f()$ is the gaussian filter and $IK()$ the analytic IK algorithm. The resulting pose $\overrightarrow{\theta(k)}$ is analyzed in order to determine if it corresponds to a valid and natural body configuration. The system considers two limitations: (i) a valid pose must respect joint limits; and (ii) a valid pose cannot produce a collision between different body items. If the system detects an incorrect position for any of the arms, it looks for alternative poses for this arm (i.e. different arm configurations). These alternatives will try to preserve hand positions, but will move the elbow to search for the most similar valid arm pose. The joint angles corresponding to the final valid pose are returned. See previous work for further details about this method [6].

The analytic nature of this method allows to obtain the joint angles of the perceived human at a high frame rate. These obtained angles $\overrightarrow{\theta(k)}$ will be the output of the HMC behaviour, and can be used to make a 3D avatar imitate the human pose. In the system described in this paper, however, an additional step is required, as depicted in Fig. 4, that translates the movements from the human motion space to the robot motion space. These spaces may be different, due to the physical differences between the human and the robot. Thus, one-to-one mapping is not adequate. Instead, the retargeting module depicted in Fig. 4 performs a combined retargeting procedure, in which two different strategies are used. The first of these strategies tries to preserve the relative positions of the end effectors with respect to the head, and is used for *location* movements, e.g. pointing. The other strategy tries to preserve the trajectories of the joint angles, and is used for *configured* movements, e.g. waving a hand to mean 'hello' [41]. As it has been mentioned above, the result of the retargeting module is a combination of both strategies, in which the weight of each factor depends on the amplitude of performed motion. A deeper explanation of this combined retargeting strategy can be found in [8].

Finally, it must be pointed out that the behaviour is reinitialised when a tracking problem arises at the semi-attentive stage and object records associated to face or hands contain invalid values. As long as the face is not lost, the behaviour can recover from these situations, as lost hands can be looked for again inside the silhouette in further frames. However, if the face is lost, it will be necessary to look for a new face, which could correspond to a different human. This mismatching problem can be avoided if the face recognition behaviour is conducted to associate each face record with a label. This strategy ensures that tracking does not continue until the right person is found.

4.3. Scene exploration behaviour

Reliable navigation is an essential component of a social robot. In order to perform this task correctly, the robot will typically need to represent the information perceived by external sensors into a navigation map. One popular choice is to build this map with distinguished natural landmarks that the robot can acquire from the environment without human supervision. Recognizable landmarks are essential since they will be used as reference marks to identify locations in the environment. In the past, a variety of approaches for feature-based mobile robot localization and naviga-

tion has been developed. A significant number of these approaches use range sensors to detect distinguished landmarks. However, although these approaches can robustly address the landmark detection problem, they become less robust to achieve the landmark description one. An alternative to the active ranging devices are vision systems. These systems are passive and of high resolution, and they provide a huge amount of information (color, texture or shape) which allow disambiguating landmarks for subsequent data association purposes. In this paper, we only address the visual landmark detection problem. Thus, perceived 'proto-objects' of data-dependent shape which satisfy certain constraints will serve as natural landmarks.

As it was aforementioned, the preattentive stage provides a contrast-based saliency map. The visual landmark detection behaviour selects, among the set of most salient 'proto-objects' of this map, those which satisfy certain conditions. The key idea is to use as landmarks rectangular shaped objects or quasi-rectangular shaped objects without holes. In this way, we try to avoid the selection of segmentation artifacts, assuming that a rectangular object has less probability to be a segmentation error than a sparse region with a complex shape. Selected 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 objects 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 non-planar surfaces. Finally, The selected objects must also exhibit a relatively high contrast with respect to its surroundings. Figure 7 shows several set of extracted landmarks.

A detailed implementation of this landmark detection method can be found in [47].

5. Experimental results

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 [8]. The robot will be provided with two arms that has four degrees of freedom (DOF), plus 1 DOF in the waist and 3 DOF in the head. 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 color 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.

Table 1
Tracking errors averaged over 5300 frames

Marker	Left shoulder	Left elbow	Left hand
Mean error (cm)	5.50	11.78	11.47
Standard deviation (cm)	3.25	7.03	6.43
Marker	Right shoulder	Right elbow	Right hand
Mean error (cm)	6.23	12.50	11.33
Standard deviation (cm)	4.53	6.51	7.59
Marker	Left head	Abdomen	Right head
Mean error (cm)	6.78	7.76	6.47
Standard deviation (cm)	4.20	1.05	5.25

5.1. Face recognition and human motion capture

Before running these behaviours over the robotic platform, a quantitative evaluation of the HMC behaviour was performed by comparing its results against ground truth obtained using a Codamotion CX1 motion capture system based on active markers.[1] This comparison is achieved for different performers by using the same setup detailed in [7]. Table 1 shows errors associated to different body parts. It can be appreciated that the elbows are affected by higher errors as their pose is not perceived, but estimated. Hand movements also accumulate a higher deviation error. This error is partially produced by the hands moving usually faster than other upper-body parts, and thus being more sensitive to tracking errors. But the main cause of these errors is the proximity to image borders. Hand movements tend to spread near these borders, where experimental results show that position errors are more significative. Thus, although the calibration software provided by Videre design[2] allows taking into account the radial and tangential distortions of the lens, obtained results show that these distortions still affect to the correct estimation of the 3D position

[1] http://www.codamotion.com/

[2] http://www.videredesign.com/

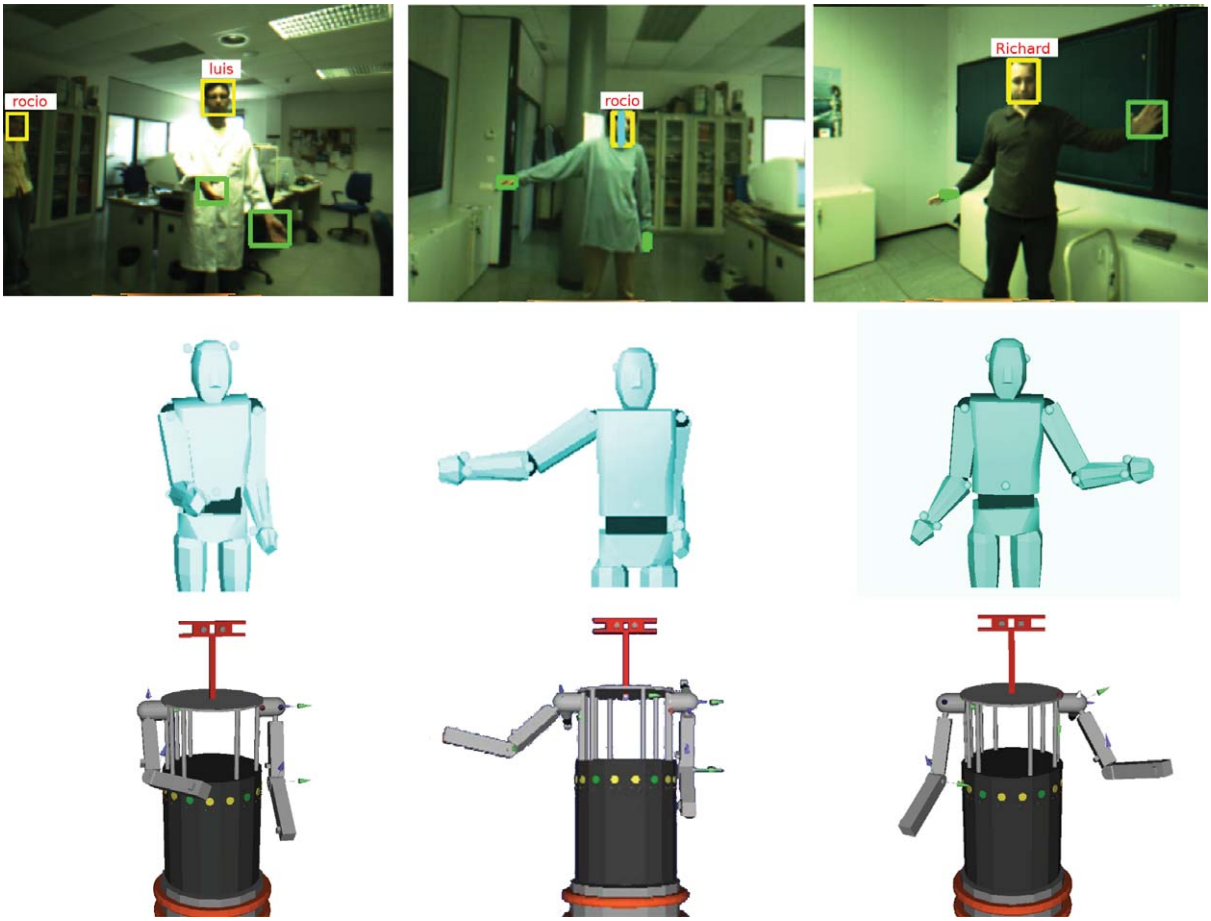

Fig. 6. Frames captured from sequences used to test the face recognition and HMC behaviours.

of perceived points. Consequently hand positions, that tend to approach image borders, accumulate a higher error.

The previous quantitative evaluation shows that results of the face recognition behaviour are similar to those obtained by [9], thus as it is concluded in that paper and our experiments confirm, faces can be recognized *on-line*, once a fast learning process has been performed. On the other hand, results also show that, although fine details such as finger movements will not be captured, the proposed HMC behaviour is adequate to obtain upper-body gestures, and retarget them to the robot motion space. Figure 6 shows results obtained when both face recognition and HMC behaviours are executed on the previously described robotic system. As depicted, the face recognition behaviour is able to detect and recognize human faces in the field of view of the robot. Then, the HMC behaviour is executed and capture the upper-body motion of the closer human at about 15 frames per second.[3] This motion is translated to the robot using the previously described retargeting strategy, that is deeply evaluated in [8]. As Fig. 6 depicts, the performer is not wearing special markers nor specific clothes. Finally, it can be seen in the last depicted frame (Fig. 6c) how the recovered 3D position for a hand may deviate from desired pose as the hand approach the image border.

5.2. *Scene exploration*

The robot was driven through different environments while capturing real-life stereo images to test the validity of the landmark detector. Figure 7 shows some

[3] Some videos showing the performance of the tracking process can be found at http://www.grupoisis.uma.es/

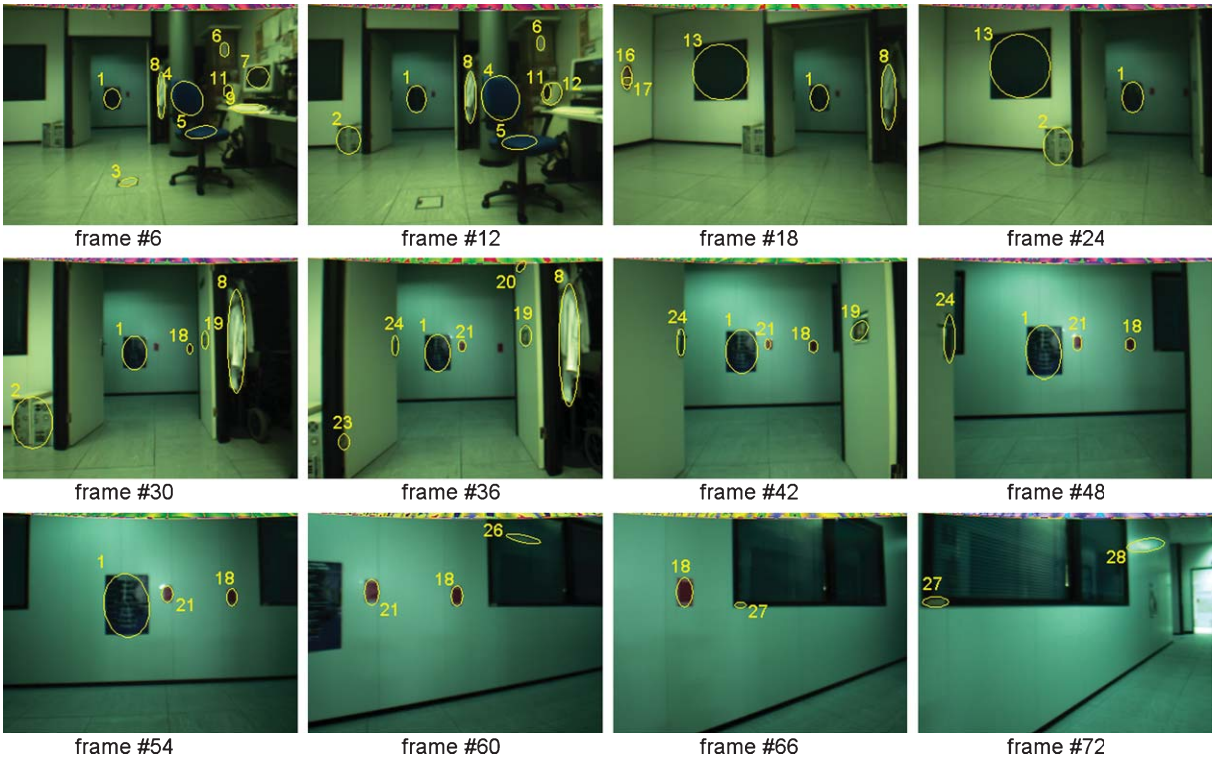

Fig. 7. Visual landmarks detected by the proposed perception system. Landmarks are represented as ellipses, and the correspondence between landmarks in different frames with the corresponding index.

frames of a trajectory in an indoor environment. The visual landmark detection behaviour provides landmarks to a simultaneous localization and mapping (SLAM) algorithm, which is the responsible of create and update the navigation map. In the figure, only landmarks which have been finally included in the map are shown. It can be seen that some erroneous landmarks (such as landmark 28 in frame #72) may appear due to reflections or noise. While it is not possible to filter these landmarks at this level, they could be removed in a posterior, higher level layer in which an object recognition algorithm would be used to decide whether a detected landmark is a meaningful object or not.

To quantitatively check the performance of our detector and compare our method to other similar approaches, images, Matlab code to carry out the performance tests, and binaries of other detectors have been downloaded from http://www.robots.ox.ac.uk/~vgg/research/affine. The database is composed by eight different image sets that represent five changes in imaging conditions (viewpoint changes, scaling, image blur, jpeg compression and illumination changes). To deal with this data set, the disparity information must not be taken into account at the preattentive stage.

Image sets can be grouped into two different scene types. Thus, one scene type contains homogeneous regions which present distinctive boundaries (structured scenes), and the other contains repeated textures of different forms (textured scenes). Our approach is based on structure cues in images. Thus, although it can work in unstructured environments, it exhibits a superior performance on structured scenes. To evaluate the detectors, we use the repeatability score [35]. The objective of this test is to measure how many of the detected regions are found in images under different transformations, relative to the lowest total number of regions detected (where only the part of the image that is visible in both images is taken into account). In all cases, the ground truth is provided by mapping the regions detected on the images in a set to the image of highest quality of this set (reference image) using homographies. The measure of repeatability is the relative amount of overlap between regions detected in the reference image and in the other image. This region is

Table 2
Repeatability scores (%) for two databases (GRAF and BOAT sequences) (see Mikolajczyk et al. (2006))

	BOAT sequence					GRAF sequence				
	Scale changes					Viewpoint angle				
Detector	1.1	1.35	1.9	2.35	2.8	20	30	40	50	60
MSER	65	63	55	50	22	78	69	64	60	46
IBR	49	51	41	32	20	61	54	45	39	29
Fast-Hessian	70	68	68	69	41	68	55	27	0	0
Proposed	71	68	65	57	35	73	69	67	63	57

projected onto the reference image using the homography relating the images. It must be noted that the output for our detector is a set of arbitrarily shaped regions. However, for the purpose of the comparisons using the Matlab code previously mentioned, the output region of all detectors are represented by an ellipse. In our case, ellipses which have the same first and second moments as the detected regions are used to approximate them. The proposed detector is compared to the *maximally stable extremal region* detector (MSER) [34], the *intensity extrema-based region* detector (IBR) [46] and the Fast-Hessian [10]. For all experiments, the default parameters given by the authors are used for each detector. The repeatability for two sets of images are illustrated in Table 2. Similar results are obtained for the rest of sequences. These results show that the proposed detector ranks similar to the rest of approaches when it deals with structured images. In these images, only few regions are detected and the thresholds can be set very sharply, resulting in very stable regions.

6. Conclusions and future work

This paper has presented a visual perception system for a social robot which main component is an attentional mechanism. This attentional mechanism integrates bottom-up and top-down processings to select the most relevant information from the broad visual input depending not only on the sensed features but also in the currently executed task. This model employs two task-independent stages: the preattentive and the semiattentive stages, and a task-dependent stage or attentive stage. The preattentive stage divides the visual scene into a set of 'proto-objects'. This allows the proposed attentional model to direct the attention on candidate to real objects, similarly to the behaviour observed in humans. 'proto-objects' are stored at the semi-attentive stage as hierarchical templates. This representation is used by the fast tracking algorithm that implements the 'inhibition of return' at this stage. The attentive stage controls the field of attention following several behaviours. Specifically, we have incorporated and tested three different behaviours. The first two behaviours -face recognition and upper-body human motion capture- provide the robot interaction abilities. The third behaviour -scene exploration- allows to autonomously acquire visual landmarks for mobile robot simultaneous localization and mapping.

Finally, in order to facilitate the interaction with people, the social robot should be able to navigate and to notice, at the same time, if there are people in the scene that are interested in interat with it. To do that, the different behaviours must be capable to run simultaneously, being necessary to implement as future work a mechanism to control this process. On the other hand, and also to facilitate the interaction with people, the information generated by the attention stage should include semantic information and not only spatial relations. Future work will be focused on integrating the robot perception abilities with human-robot interaction processes which will allow the robot to annotate its internal representations with semantic information in a supervised way. The importance of the semantic information has been largely pointed out in the robotic literature. Thus, it can be used to reason about the functionalities of objects and environments, or to provide additional input to the navigation and localization modules. In any case, it is fundamental to allow the robot to communicate with people using a common set of terms and concepts [21].

Acknowledgements

This work has been partially granted by the Spanish Ministerio de Educación y Ciencia (MEC) and FEDER funds, Project n. TIN2008-06196 and by the Junta de Andalucía, Project n. P07-TIC-03106. The authors would like to thank Dr. Adrian Hilton for allow us using the CODA Motion system available at the CVSSP (University of Surrey).

References

[1] P. Azad, A. Ude, T. Asfour and R. Dillmann, Stereo-based markerless human motion capture for humanoid robot systems, *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, IEEE, Rome, Italy, 2007, 10–14.

[2] M.Z. Aziz and B. Mertsching, Color saliency and inhibition using static and dynamic scenes in region based visual attention, WAPCV 2007, LNAI, L. Paletta and E. Rome, eds, (Springer, Heidelberg) **4840** (2007), 234–250.

[3] M.Z. Aziz and B. Mertsching, Visual Attention in 3D Space, *Proceedings of the 1st International Conference on Informatics in Control*, Automation and Robotics, 2008; Springer, Milan, Italy, 2009.

[4] G. Backer, B. Mertsching and M. Bollmann, Data- and model-driven gaze control for an active-vision system, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **23** (2001), 1415–1429.

[5] G. Backer and B. Mertsching, Two selection stages provide efficient object-based attentional control for dynamic vision, *International Workshop on Attention and Performance in Computer Vision (WAPCV 2003)*; 2003 April 3; Springer, Heidelberg, Graz, Austria, 2003.

[6] J.P. Bandera, R. Marfil, L. Molina-Tanco, J.A. Rodríguez, A. Bandera and F. Sandoval, Robot learning by active imitation, Humanoid Robots: Human-like Machines, M. Hackel, ed, (ARS), 2007.

[7] J.P. Bandera, R. Marfil, J.A. Rodríguez, L. Molina-Tanco and F. Sandoval, A novel hybrid approach to upper-body human motion capture, *Proceedings of the 14th Mediterranean Electrotechnical Conference*, IEEE, Ajaccio, France, 2008.

[8] J.P. Bandera, R. Marfil, R. López, J.C. del Toro, A. Palomino and F. Sandoval, Retargeting system for a social robot imitation interface, *Proceedings of the 11th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines (CLAWAR 2008)*; 2008 September; World Scientific Publishing, Coimbra, Portugal, 2008.

[9] J. Barreto, P. Menezes and J. Dias, Human-robot interaction based on Haar-like features and eigenfaces, *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2004)*; 26, May 1, IEEE, New Orleans, USA, 2004

[10] H. Bay, T. Tuytelaars and L. Van Gool, SURF: Speeded Up Robust Features, Computer Vision – ECCV (2006), 404–417.

[11] M. Begum, G.K.I. Mann, R. Gosine and F. Karray, Object- and space-based visual attention: An integrated framework for autonomous robots, *Procedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008 September 22–26; IEEE, Nice, France, 2008.

[12] I. Bogdanova, A. Bur and H. Hugli, Visual Attention on the Sphere, *IEEE Transactions on Image Processing* **17**(11) (2008), 2000–2014.

[13] C. Breazeal, A. Brooks, J. Gray, M. Hancher, J. McBean, D. Stiehl and J. Strickon, Interactive robot theatre, *Communications of the ACM* **46**(7) (2003) 76–84.

[14] R. Contini, Body Segment Parameters. Part II, *Artificial Limbs* **16**(1) (1972) 1–19.

[15] A. Cruz, J.P. Bandera and F. Sandoval, Torso pose estimator for a robot imitation framework, *Proceedings of the 12th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines (CLAWAR 2009)*; 2009 September 9–11, World Scientific Publishing, Istanbul, Turkey, 2009.

[16] A. Dankers, N. Barnes and A. Zelinsky, A reactive vision system: active-dynamic saliency, *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS'07)*; 2007 March 21–24, Applied Computer Science Group, Bielefeld University, Bielefeld, Germany, 2007.

[17] K. Dautenhahn and C. Nehaniv, Imitation in animals and artifacts, MIT Press, Cambridge, 2002.

[18] C.W. Eriksen and Y.Y. Yenh, Allocation of attention in the visual field, *Journal of Experimental Psychology: Human Perception and Performance* **11**(5) (1985), 583-597.

[19] T. Fong, I. Nourbakshsh and K. Dautenhahn, A survey of social robots, *Robotics and Autonomous Systems* **42** (2002), 143–166.

[20] S. Frintrop, G. Backer and E. Rome, Goal-directed search with a top-down modulated computational attention system. DAGM 2005, LNCS 3663, W.G. Kropatsch, R. Sablatnig and A. Hanbury, eds, (Springer, Heidelberg), 2005.

[21] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J.A. Fernández-Madrigal and J. González, Multi-Hierarchical semantic maps for mobile robotics, *Proceedings of the IEEE/RSJ Int. Conf. on Intell. Robots and Systems*, IEEE, Edmonton, Canada, 2005.

[22] L. Itti, U. Koch and E. Niebur, A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **20** (1998), 1254–1259.

[23] L. Itti, Real-time high-performance attention focusing in outdoors color video streams, *Proceedings of the SPIE Human Vision and Electronic Imaging (HVEI'02)*; 2002 January 21–24; The International Society for Optical Engineering, San Jose, USA, 2002.

[24] C. Koch and S. Ullman, Shifts selective visual attention: Towards the underlying neural circuitry, *Human Neurobiology* **4** (1985), 219-227.

[25] N. Kojo, T. Inamura, K. Okada and M. Inaba, Gesture Recognition for Humanoids using Proto-symbol Space, *Proceedings of the 6th IEEE-RAS International Conference on Humanoid Robots (Humanoids'06)*; 2006 December; IEEE, Genoa, Italy, 2006.

[26] M. Maire, P. Arbeláez, C. Fowlkes and J. Malik, Using contours to detect and localize junctions in natural images, *Proceedings of the Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2008)*; 2008 June 24–26, Anchorage, Alaska, 2008.

[27] A. Maki, P. Nordlund and J.O. Eklundh, Attentional scene segmentation: integrating depth and motion. *Computer Vision and Image Understanding* **78**(3) (2000), 351–373.

[28] R. Marfil, L. Molina-Tanco, A. Bandera, J.A. Rodríguez and F. Sandoval, Pyramid segmentation algorithms revisited, *Pattern Recognition* **39**(8) (2006), 1430–1451.

[29] R. Marfil, L. Molina-Tanco, J.A. Rodríguez and F. Sandoval, Real-time object tracking using bounded irregular pyramids, *Pattern Recognition Letters* **28** (2007), 985–1001.

[30] R. Marfil, A. Bandera and F. Sandoval, Perception-based image segmentation using the Bounded Irregular Pyramid, *LNCS* **4713** (2007) 244–23.

[31] R. Marfil, L. Molina-Tanco, A. Bandera and F. Sandoval, The construction of bounded irregular pyramids using a union-find decimation process. GbRPR 2007, *LNCS* **4538** (2007), 307–318.

[32] R. Marfil, A. Bandera, J.A. Rodríguez and F. Sandoval, A novel hierarchical framework for object-based visual attention. Attention in cognitive systems. *LNCS* **5395** (2008), 27–40.

[33] D. Martin, C. Fowlkes, D. Tal and J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, *Proceedings of the Eighth IEEE Int. Conf. on Computer Vision (ICCV 2001)*; 2001 July 7–14, IEEE, Vancouver, Canada, 2001.

[34] J. Matas, O. Chum, M. Urban and T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, *Proceedings of the British Machine Vision Conference*, 2002 September 2–5, 2002.

[35] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool, A comparison of affine region detectors, *Int. Journal Computer Vision* **65** (2006), 43–72.

[36] T.B. Moeslund, A. Hilton and V. Krüger, A survey of advances in vision-based human motion capture and analysis, *Computer Vision and Image Understanding* **104** (2006), 90–126.

[37] C.R. Olson, Object-based vision and attention in primates, *Curr. Opin. Neurobiol* **11**(2) (2001) 171–179.

[38] F. Orabona, G. Metta and G. Sandini, A proto-object based visual attention model. WAPCV 2007, *LNAI* **4840** (2007), 198–215.

[39] Z.W. Pylyshyn, Visual indexes, preconceptual objects, and situated vision, *Cognition* **80**(1–2) (2001), 127–158.

[40] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor and R. Pfeifer, Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub, *Proceedings of the IEEE International Conference on Robotics and Automation*; 2008 May 19–23, IEEE, Pasadena, California, 2008.

[41] M.M. Smyth and L.R. Pendleton, Space and movement in working memory, *The Quarterly Journal of Experimental Psychology Section A* **42**(2) (1990), 291–304.

[42] Y. Sun and R.B. Fisher, Object-based visual attention for computer vision, *Artificial Intelligence* **146**(1) (2003), 77–123.

[43] J. Terrillon and S. Akamatsu, Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images, *Proceedings of the 12th Conf. on Vision Interface*, IEEE Computer Society, 1999.

[44] S.P. Tipper, Object-centred inhibition of return of visual attention, *Quarterly Journal of Experimental Psychology* **43** (1991), 289–298.

[45] A.M. Treisman and G. Gelade, A feature integration theory of attention, *Cognitive Psychology* **12**(1) (1980), 97–136.

[46] T. Tuytelaars and L. Van Gool, Matching widely separated views based on affine invariant regions, *Int. Journal on Computer Vision* **59**(1) (2004), 61-85.

[47] R. Vázquez-Martín, onLine Environment Segmentation based on Spectral Mapping (LESS-Mapping) [*PhD thesis*]. University of Málaga, [Málaga (Spain)], 2009.

[48] P. Viola and M. Jones, Robust real-time face detection, *Int. Journal of Computer Vision* **57**(2) (2004), 137–154.