Towards a Spatial Model for Humanoid Social Robots*

Dario Figueira[1], Manuel Lopes[1], Rodrigo Ventura[1], and Jonas Ruesch[2]

[1] Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal
{dfigueira,macl,rodrigo.ventura}@isr.ist.utl.pt
[2] Artificial Intelligence Laboratory, Department of Informatics
University of Zurich
Switzerland
ruesch@ifi.uzh.ch

Abstract. This paper presents an approach to endow a humanoid robot with the capability of learning new objects and recognizing them in an unstructured environment. New objects are learnt, whenever an unrecognized one is found within a certain (small) distance from the robot head. Recognized objects are mapped to an ego-centric frame of reference, which together with a simple short-term memory mechanism, makes this mapping persistent. This allows the robot to be aware of their presence even if temporarily out of the field of view, thus providing a primary spatial model of the environment (as far as known objects are concerned). SIFT features are used, not only for recognizing previously learnt objects, but also to allow the robot to estimate their distance (depth perception). The humanoid platform used for the experiments was the iCub humanoid robot. This capability functions together with iCub's low-level attention system: recognized objects enact salience thus attracting the robot attention, by gazing at them, each one in turn. We claim that the presented approach is a contribution towards linking a bottom-up attention system with top-down cognitive information.

1 INTRODUCTION

Although Artificial Intelligence (AI) having accomplished notable results on many specific domains, being Kasparov's defeat to Deep Blue in 1997 one popular account of that success [1], general intelligence constitutes, still, largely an open issue [2]. This is particularly true for the case of physical agents, namely robots. Unlike symbolic environments, for which sophisticated AI techniques have been developed (reasoning, knowledge representation, and so on), physical agents operating in the real world demand several "basic" problems to be solved.

* This work was supported by the European Commission, Project IST-004370 RobotCub, and by the Portuguese Government — FCT (ISR/IST plurianual funding), and through the project BIO-LOOK, PTDC / EEA-ACR / 71032 / 2006

One such problem is *perception*, in the sense of associating raw sensory data with internal representations.

Consider, for instance, a cup that is positioned in the field of view of the robot. The presence of this object is expected to enact some kind of internal representation. Does the robot recognize this cup in particular, or as a new, unknown object? is it graspable? However, for this level of representation, the physical nature of the robot demands the designer to deal with the problem of how to bridge the gap between the pixels that correspond to the cup (and to the background), and the concept of *object*.

A variety of cognitive architectures has been proposed with the goal of answering the problem of general artificial intelligence [3]. One aspect common to virtually all cognitive architectures is the concept of *object*. Most models assume that the system is capable of perceptually segmenting the world in objects, some of which can be grasped, while others cannot, as in the previous example of the cup.

Instead of taking one particular architecture, we address the problem of robot intelligence following a bottom-up approach, *i.e.,* constructing building blocks towards cognitive behavior. This approach, being both constructionist and minimally constrained by prior assumptions, ends up being fairly agnostic in terms of the cognitive architecture where it can be part of.

This paper addresses the problem of learning new objects and representing their presence in the environment in a spatial model. This capability builds upon an existing perceptually-driven attention system. The proposed spatial model tackles two aspects: (1) the identification of the object, and (2) its position in the environment. For the first aspect, Scale Invariant Feature Transform (SIFT) [4] features are being used to learn new objects and to recognize them in the environment.

The existing attention system provides a saliency map with respect to a robot-centric coordinate system (ego-sphere) [5]. This saliency map, together with a inhibition of return mechanism (IOR), allows the robot to saccade from salient point to salient point. However, these salient points correspond to preattentive features, e.g., movement, color, and shape, that do not incorporate the concept of object.

The environment is being modeled with a saliency map [6]. Objects are recognized continuously in the camera images by the SIFT algorithm. The recognized objects in the robots field of vision are inserted into the egocentric map. Each one of these objects enacts one salient point, which attracts the robot attention. Thus, with several known objects, the robot is expected to commute automatically its attention focus from recognized object to recognized object. Moreover, this saliency persists even when the robot is not directly looking at them. Instead of the short-time memory of the previous works [5, 6], the system remembers where known objects are at longer time scales.

We also implemented an algorithm to automatically store to a database new objects as they get close to the robot. This draws from the idea of grasping an

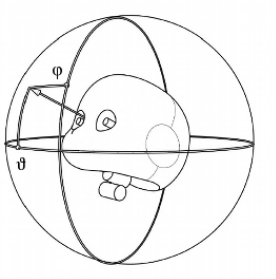


Fig. 1. Ego-sphere: a spherical map of the surroundings with a spherical coordinate system (azimuth ϑ and elevation φ.)

object for inspection. As the robot has no arms (at the time of the experimental work), this was replaced by the detection of proximity.

To do so, we compute the depth of each pair of matched SIFT features by computing the disparity between them. By using this algorithm for detecting distance, we define a new object as a cluster of SIFT features in close proximity to the cameras. This approach is robust to non-convex objects as well as objects with holes, but learns two objects shown side by side as a single object.

By integrating this capability into the existing architecture, the attention module will be able to acknowledge the saliency of known objects, because they are recognized as such. Moreover, the capability of recognizing known objects by visual features paves the way for higher level cognitive modules, such as language.

We project the surrounding space and objects into a spherical coordinate system centered in the neck of the robot, an egocentric sphere or ego-sphere, as defined in [5] (Figure 1). A spatial model for the robot is here understood as a model representing the environment surrounding it, namely the known objects, together with their relative positions to the robot.

The research described here was carried out using a humanoid robotic head, composed of an anthropomorphic head with 6 degrees of freedom, and a pair of stereo cameras with individual pan and common tilt. This head is part of the iCub humanoid robot, which has been designed as a platform for research on cognition from a developmental point of view [7]. We consider a robot centered coordinate system, specifically, a torso anchored coordinate system. The software module that resulted from this research fits well into the iCub software architecture being developed in the context of the RobotCub project[3].

The problem of automatically associating sensor data with internal symbolic representations has been formulated in a domain independent way as the symbol anchoring problem [8, 9]. These approaches assume however a dualistic view of perception and symbolic representations. Another approach, closer to representations of sensorimotor nature, concerns learning affordances of objects

[3] http://www.robotcub.org/

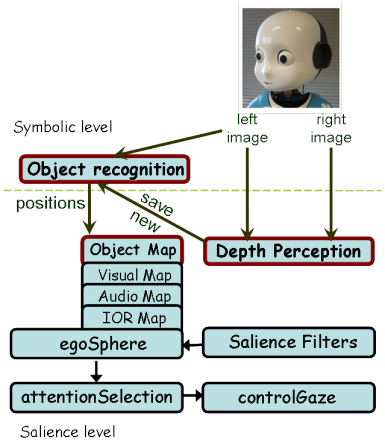


Fig. 2. System architecture, after introducing the modules presented here (dark red border): object recognition, depth perception, and a new map, the objects map.

by interaction [10, 11]. The identification of objects is often simplified by using colored objects and pre-wired recognition algorithms.

In the next section the architecture of the proposed system is described, leaving out the implementation details to be discussed in the third section. The forth section presents the experiment in which the functioning of the system is illustrated, and experimental results are presented to validate the approach. Finally, we finish with some conclusions and future work.

2 ARCHITECTURE

The architecture, displayed in Figure 2, comprises several interconnected modules, forming a sensing-deliberation-actuation chain. It is motivated on the Itti and Koch model [12] where stimuli from various sources are represented and combined into a single saliency map. Then, the point that maximizes this map is selected, as the one winning the robot attention. Finally the robot gazes towards the new selected attention point.

The ego-sphere keeps a short-memory of the previously looked upon positions, in the form of an inhibition-of-return mechanism (IOR). The IOR information reduces the salience levels of the already observed locations. The resulting behavior is the capability of the robot to fully explore its environment without being stuck on the absolute maxima of the saliency maps.

Our work adds a level of abstraction — the concept of object — to the previous architecture. To acquire this knowledge the robot has to solve two problems. *What* and *when* to learn a new object. A new object is learned when it is detected in close proximity of the robot. The object is assumed to consist in the image patch that is close to the robot. The proximity measure is defined based on the arms length distance, *i.e.,* the reachable objects. New objects are stored

together with a "label", corresponding to a number (the order of appearance). When the robot has the possibility to ask humans around him for the names of the objects it is discovering and storing, new possibilities concerning associating object representations to names arise.

3 IMPLEMENTATION

3.1 Object recognition

Many different approaches have been used in computer vision to enable recognition, for instance, eigenspace matching has been used successfully by Schiele [13], others have used Speeded Up Robust Features (SURF) [14], and many have benefited from David Lowe's Scale Invariant Feature Transform (SIFT) [4]. We have followed the latter, tackling both problems of object segmentation and recognition. We chose SIFT over eigenspace matching for reasons such as invariance to scale and excelling in cluttered or occluded environments (as long as three SIFT features are detected, the object is recognized). And while the SURF algorithm is faster and performs generally well, SIFT's recognition results are still superior [15].

SIFT [4] is an algorithm that extracts, features from an image. These features are computed from histograms of the gradients around the key-points, and are not only scale invariant features, but also invariant to affine transformations (*e.g.,* rotations invariant). Furthermore, they are robust to changes in lighting, robust to non-extreme projective transformations, robust up to 90% occlusion, and are minimally affected by noise. We use the SIFT algorithm to enable the recognition in our system because of all these powerful characteristics. However, we observed two drawbacks on its use. The first one is that it is computationally expensive, as the most efficient and freely available implementations are not able to run in real time. Due to the nature of the SIFT features, its second drawback is the inability to extract features from a texture-less object, as shown in Figure 3: few or no features, in yellow dots, are found in areas with homogeneous color, such as on the table, on the ground, or on the wall.

3.2 Depth Perception

The common way to determine depth, with two stereo cameras, is by calculating disparity. Disparity is defined as the subtraction, from the left image to the right image, of the 2D coordinates of corresponding points in image space. To calculate depth we require the knowledge of the following camera parameters: focal length f, camera baseline β, and pixel dimension γ. Also, we need to correctly match a point of the environment, seen in both stereo images, with pixel coordinates (x_1, y_1) in the first image and (x_2, y_2) in the second. The point's coordinates in the camera references are (X_1, Y_1, Z) for the first camera and (X_2, Y_2, Z) for the second. Then, we can calculate how far away the matched point is (depth Z) by

Fig. 3. Example of SIFT feature extraction; the dots (yellow) correspond to the extracted features positions.

derivation (1), and illustrated in Figure 4.

$$\begin{cases} \gamma\, x_1 = f\, \frac{X_1}{Z} \\ \gamma\, x_2 = f\, \frac{X_2}{Z} \end{cases} \Leftrightarrow Z = \frac{f\, \beta}{\gamma\, (x_1 - x_2)} \tag{1}$$

where $\beta = X_1 - X_2$.

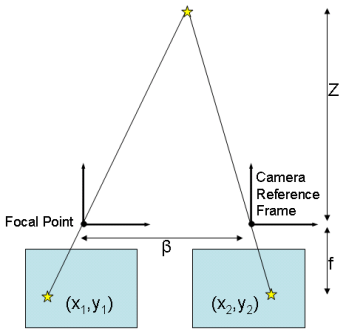


Fig. 4. Pinhole camera model used to calculate depth, f: focal distance, β: distance separating the parallel cameras, γ: pixel-to-meter ratio in the camera sensors, (x_1, y_1): pixel coordinates of point we wish to calculate depth, Z: depth.

Usual ways of match corresponding points include pixel by pixel probabilistic matching with a Bayesian formulation [16], and histogram matching of the neighborhood of the pixel [17].

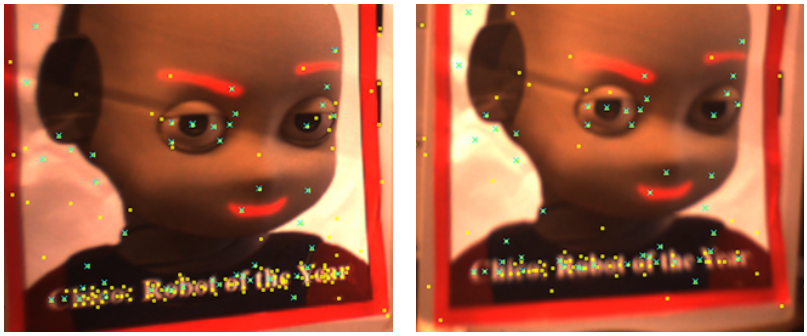
The SIFT features, with their invariance and robustness, suggest a different approach to solve the problem of matching corresponding points in stereo images. We generate a sparse disparity map by extracting the SIFT features from stereo images, and look for matches between both sets. Assuming that the robot's eyes

are roughly aligned in the horizontal (*i.e.,* misalignment of under 30 pixels) we compute the disparity between matching features from the pair of stereo images. Matches that have a high horizontal disparity are assumed to be part of an object in close proximity to the robot's face and matches with low horizontal disparity belong to the "background." Matches with high vertical disparity or negative horizontal disparity are considered outliers, and thus discarded.

Comparing the extracted features of different images in different resolutions [18], a threshold for the horizontal disparity T_h was found empirically to be the width of the image divided by 6.4. Moreover, the vertical threshold T_v to determine outliers was also empirically found to be the height of the image divided by 16.

If the matches between detected features are close enough (each match having its horizontal disparity greater than the threshold), the group is stored in the database as a new object. Only the features that are correctly matched between the two stereo images with high horizontal disparities are stored, because only these features are believed to belong to the close object. For instance, the features from the background being seen by a hole in the object are then automatically ignored.

Figure 5(a) and Figure 5(b) exemplify in blue crosses the features that are correctly matched between the two stereo images as being the same, and therefore stored to the database as a new object (if not recognized as part of an already known object).



(a) Left image (b) Right image

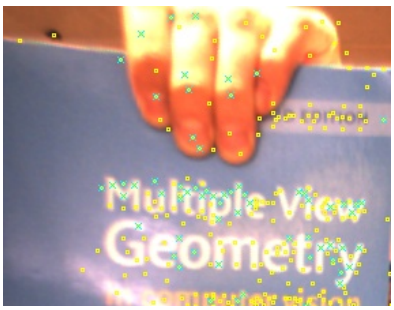
Fig. 5. Matching SIFT features in a pair stereo images: features in dots (yellow); matched features in crosses (blue). Images cropped for clarity.

3.3 Recognition

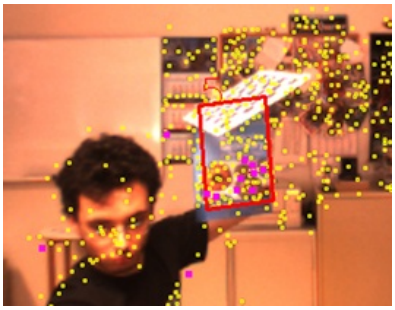
To decide upon the presence of an object in the image, SIFT relies on a voting mechanism that is implemented by a Hough transform. Defining pose as the

position, rotation and scale of an object, each match votes on an object-pose pair in the image. The Hough transform is computed to identify clusters of matches belonging to the same object. Finally, a verification through least-mean-squares is conducted for consistent pose parameters along all matches (verifying if the matches found have correct relative positions).

After experimenting with several objects, having the robot store them to the database and then holding them farther and farther away, the algorithm was able to recognize them until roughly two meters away, when the number of extracted features declines significantly. Of the many features stored in the database and shown in blue crosses in Figure 6(a), only the few extracted ones, depicted in purple filled squares, are needed to recognize the object in Figure 6(b).



(a) Object saved to database; SIFT features in dots (yellow), and SIFT features saved to the database in crosses (blue).



(b) Marked rectangle (in red): recognized object, SIFT features in dots (yellow), and SIFT features matched with the database in squares (purple).

Fig. 6. Recognition of saved object in the environment. Images cropped for clarity.

3.4 Database and Mapping

New objects are stored into a database, associating object identifiers (labels) to sets of SIFT features. When an object is recognized in the environment, its position is mapped into the ego-sphere [5]. Object representations are stored in the database (long-term memory), while their positions, whenever recognized by the robot, are represented solely in the ego-sphere (short-term memory, forming the robot's spatial model).

The egocentric saliency map used for attention selection is obtained from the composition of several specialized maps: a visual map (M_{vis}), containing saliency information extracted from visual features (e.g., motion, color), and an auditory map (M_{aud}), obtained from sound stimuli captured by the robot's microphones [5]. These maps cover the entire space surrounding the robot with a spherical coordinate system (azimuth $\vartheta \in [-180\,°; 180\,°]$ and elevation $\varphi \in$

$[-90\,°; 90\,°]$). The saliency information stored in these maps is continuously decayed ($M_{vis}(k+1) = d_{vis}\, M_{vis}(k)$, $M_{aud}(k+1) = d_{aud}\, M_{aud}(k)$), according to a forgetting factor ($d_{vis} = d_{aud} = 0.95$ in the experiments). This factor together with a maximum frame-rate of 20 FPS, yields a half-life of less than a second, 14 frames.

In order to integrate the system described in this paper with the attention selection mechanism, the recognized objects are projected onto a third map (an object map M_{obj}). This map, combined with the other two, contributes for the ego-centric saliency map: $M_{ego} = max(M_{vis}, M_{aud}, M_{obj})$. As the others, this map is also subject to a continuous decay of its information, albeit with a much longer forgetting factor ($M_{obj}(k+1) = d_{obj}\, M_{obj}(k)$, where $d_{obj} = 0.9995$ in the experiments). How long should the robot remember where objects of interest were? How long before such information is unreliable? Those are not trivial questions to answer, at least with contextual knowledge. Therefore, to fulfill the practical goal of this work, of enabling the robot to switch its attention focus from recognized object to recognized object, even when such objects are not continuously in the robots field of view, this simple decaying memory with such a forgetting factor, that gives an half-life of little over one minute, seems sufficient.

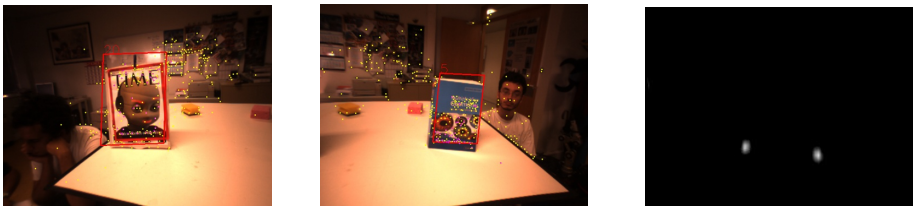
To verify the repeatability of the mapping of an object position from coordinates (x, y) of in the image frame to the corresponding coordinates (ϑ, φ) in the ego-sphere frame the following experiment were conducted: An object was left on the table in front of the robot, while the robot's head was slowly turned. We concluded that, when the object is visible, it is repeatedly mapped to the same location with an error under one degree elevation and two degrees azimuth. When on the verge of leaving the image, the error in mapping jumps up to two degrees elevation and four degrees azimuth.

The objects are mapped into the ego-sphere as gaussian peaks in salience. To account for the mapping uncertainty, the gaussian parameters used were $\sigma_\vartheta = 30$ and $\sigma_\varphi = 15$.

4 RESULTS

To validate the approach, several experiments were conducted. In all of them, the robot operates autonomously, meaning that object learning was triggered by its proximity to the robot, and the successful recognition of objects was verified by the robot gaze. In all these experiments, saliency depends only on known objects, *i.e.,* the other saliency maps in figure 2 were disabled.

In the first experiment, two previously learnt objects were shown to the robot, both initially visible, but sufficiently apart so that one of them is not visible while the other is being gazed at. In this experiment, the robot was able to successfully switch its gaze towards each one of the objects, pausing to gaze at each object in turn [18]. Persistence of the objects positions in the spatial model is necessary for this. Figure 7 shows the objects as seen by the robot, together with the resulting saliency map.



(a) First object (b) Second object (c) Saliency map

Fig. 7. Recognizing and gazing at objects in the environment: camera views of each one of the objects in (a) and (b) individually gazed at, and the saliency map (c) showing the positions of both objects in the ego-sphere frame of reference.

The second experiment aimed at evaluating the behavior of the system with both unknown and known objects. First, an unknown object was shown, which did not attract the robot gaze. This object was then shown close to the robot, triggering its acquisition. Then, this object returned to its previous position, after which the robot was able to recognize it, gazing at it. The procedure was repeated for a second, and for a third object. In all of these steps, the robot ignored the new introduced object before learning it, and gazing at it afterwards. In the end, the robot shared its time among gazing at each one of the objects.

Note that the reported experiments are robust to dynamic backgrounds and non-uniform illumination of the scene, as they were performed in a busy lab, without any special preparation. This robustness is mostly due to the choice of the SIFT features for object learning and recognition.

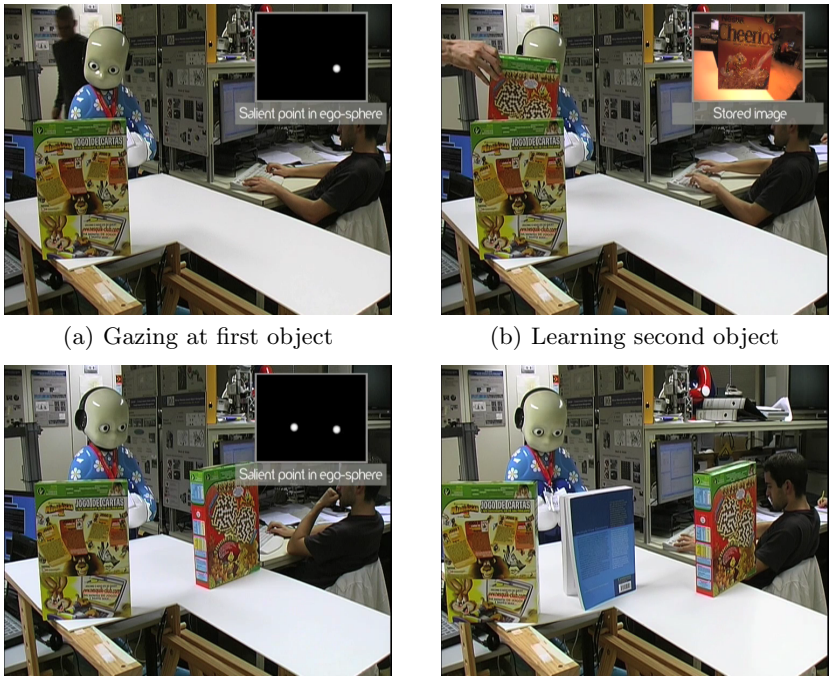
5 CONCLUSIONS AND FUTURE WORK

The work presented here aimed at the implementation of a spatial model of the space surrounding a humanoid robot, including the salient objects which the robot encounters in its explorations. This model is used to commute the robot's attention focus automatically between objects, while not being dependent on the robot field of vision, nor on the objects visibility conditions.

To this end, we mapped recognized objects by introducing salience peaks on the ego-sphere [5]. The robot can now explore its environment based on low-level saliency but also on high-level information (objects).

With this long-term object memory implemented, the goal of making this spatial model non-dependent on the robot's field of vision was achieved. As depicted in the results, the robot returns its focus to previously observed objects that were at the moment not in its field of view.

Real-time operation of the implemented system is hindered by the fact that the computation of the SIFT features is computationally demanding. In the future, we expect to improve the object recognition by introducing other kinds of features, not only to address this performance issue, but also to be able to



(a) Gazing at first object (b) Learning second object

(c) Two known objects (d) Three known objects

Fig. 8. Setup for the second experiment (see text), showing (a) the robot gazing at the first object shortly after acquisition, with the saliency map also shown, (b) learning the second object, (c) two known objects, together with the saliency map, and (d) three known objects.

recognize texture-less regions in objects, as SIFT features perform poorly on regions of this nature.

At the time of writing, the iCub platform is already equipped with arms, thus opening new possibilities in terms of integrating the module here presented with grasping behaviors of the robot. Manipulation of objects by the robot also raises interesting possibilities of combining affordances with feature-based object recognition.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of Alexandre Bernardino's reviews, comments and insights in the matter of depth perception.

References

1. Hamilton, C.M., Hedberg, S.: Modern masters of an ancient game. AI magazine **18**(4) (Winter 1997) 137–144 AAAI.

2. Nilsson, N.J.: Human-level artificial intelligence? be serious! AI magazine **26**(4) (Winter 2005) 68–75
3. Vernon, D., Metta, G., Sandini, G.: A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. IEEE Transactions on Evolutionary Computation **11**(2) (April 2007) 151–180
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (November 2004) 91–110
5. Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., Pfeifer, R.: Multimodal saliency-based bottom-up attention, a framework for the humanoid robot icub. IEEE International Conference on Robotics and Automation Pasadena, CA, USA, May, 2008 (May 2008)
6. Dankers, A., Barnes, N., Zelinsky, A.: A reactive vision system: Active-dynamic saliency. Proc. of the 5th International Conference on Computer Vision Systems, ICVS, Bielefeld, Germany (March 2007)
7. Metta, G., Sandini, G., Vernon, D., Caldwell, D., Tsagarakis, N., Beira, R., Santos-Victor, J., Ijspeert, A., Righetti, L., Cappiello, G., Stellin, G., Becchi, F.: The RobotCub project: an open framework for research in embodied cognition. In: Proceedings of the 2005 IEEE-RAS International Conference on Humanoid Robots. (2005)
8. Coradeschi, S., Saffiotti, A.: Perceptual anchoring of symbols for action. In: Proceedings of the 17th International Conference on Artificial Intelligence (IJCAI-01), Seattle, WA (2001) 407–412
9. Chella, A., Coradeschi, S., Frixione, M., Safotti, A.: Perceptual anchoring via conceptual spaces. In: Proceedings of the AAAI-04 Workshop on Anchoring Symbols to Sensor Data, AAAI Press (2004)
10. Stoytchev, A.: Toward learning the binding affordances of objects: A behavior-grounded approach. In: Proceedings of AAAI Symposium on Developmental Robotics, Stanford University (2005) 17–22
11. Montesano, L., Lopes, M., Bernardino, A., Santos-Victor, J.: Learning object affordances: From sensory motor maps to imitation. IEEE Transactions on Robotics **24**(1) (February 2008) 15–26
12. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research **40**(10-12) (May 2000) 1489–1506
13. Schiele, B., Crowley, J.L.: Object recognition using multidimensional receptive field histograms. In: European Conference on Computer Vision. (1996) I:610–619
14. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. Computer Vision (ECCV) **3951** (2006) 404–417
15. Bauer, J., Sunderhauf, N., Protzel, P.: Comparing several implementations of two recently published feature detectors. In: International Conference on Intelligent and Autonomous Systems (IAV), Toulouse, France (2007)
16. Bernardino, A., Santos-Victor, J.: A binocular stereo algorithm for log-polar foveated systems. Biological Motivated Computer Vision, Tuebingen, Germany (November 2002)
17. Prazdny, K.: Detection of binocular disparities. Biological Cybernetics **52**(2) (June 1985) 93–99
18. Figueira, D., Lopes, M., Ventura, R., Ruesch, J.: From pixels to objects: Enabling a spatial model for humanoid social robots. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA-09), Kobe, Japan (May 2009)