CuriousMind photographer: distract the robot from its initial task

Vincent Courboulay, Matei Mancas

ICST Transactions Preprint

Inciting robots to act out of curiosity

Vincent Courboulay[1,*], Matei Mancas[2]

[1]L3I, University of La Rochelle, Av Michel Crepeau, 17042 La Rochelle, France.
[2]TCTS Lab, University of Mons (UMONS) 20, Place du Parc, 7000, Mons, Belgium

Abstract

There is no doubt that robots are our future, but to be realistic they have to develop competences and abilities to interact with us. This paper introduces an attentive computational model for robots. Actually, attention if the first step to interaction. We propose to enhance and implement an existing real time computational model. Intensity, color and orientation are usually used but we have added information related to depth and isolation. We have built a robotic system based on Lego Mindstorm and Kinect, that is able to take a picture of the most interesting part of the scene.

Keywords: Attentionnal system, robotic implementation, 3D saliency.

1. Introduction

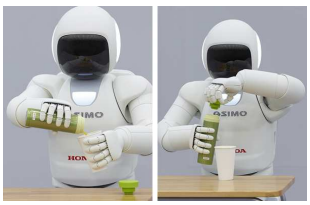
1.1. Context

Robots are our future. They will help us with all the boring daily tasks: housekeeping, shopping, classification ... We will have many interactions with intelligent robots. To do this they need to fit into our lives with comprehensive abilities: vision, grasping, motion, etc. For us human beings all of these capabilities are often conditioned by our ability to pay attention to something (person, object, word, etc). If we cannot pay attention to the world around us we can neither anticipate dangers, nor share with others. Visual attention is, by the way, an important phenomenon to be abble to understand our environment. It corresponds to the mechanisms that enable us to select visual information in order to process some clues in particular. While machine vision systems are becoming increasingly powerful, in most regards they are still far inferior to their biological counterparts. Robots are not a Darwinian evolutionary system, thus this ability will not emerge *ex nihilo*. Attention is important for robots mainly because of two reasons:

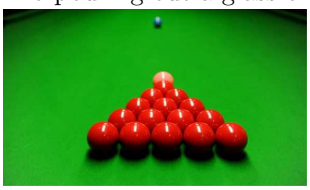
- Attention as a functional objective;
- Attention as a consequence of our limited abilities to process information.

The first point considers that our processing capabilities are unlimited. For proponents of this theory ([1], [13], [21]), attention would not be a filter for our limited brain capacities, but would be a filter for our limited capacities of action. Motor skills are limited by morphology, for example hands can only handle one (or two) objects simultaneously (cf Figure 1 (1)). Thus, action capacities are limited and require the collecting of a selection of information in order to treat it accurately.

The second theory considers irrelevant messages are filtered out before the stimulus information is processed for meaning. In other words, if our brain were bigger and/or more powerful, we would not need attentional mechanisms [2]. In this context, attention selects some information in order not to overload our cognitive system. This is also the basic premise of a large number of computational models of visual attention [10] [15] [7].



(1) Asimo pouring out a glass of water



(2) Isolated snooker blue ball

Figure 1. Illustrations

*Corresponding author. Email: vincent.courboulay@univ-lr.fr

1.2. Hypothesis

The objective of this article is to propose an attentive computational model for robots. This model is an enhancement of [10] and [4]. The main difference between the above mentioned models and those that have to be implemented on a robot mainly relies on the presence of spatial information. We propose to integrate two new conspicuity maps:

- one for the depth,
- one for the isolation.

The depth map helps promote the nearest elements. The isolation map brings out an element, even banal or diffuse, but clearly separated from the rest of its surroundings (cf Figure 1 (2)).

In the following section we describe a few computational models of attention as well as our contributions concerning a model of attention for a robotic system. In section 3 we describe how we have integrated our model in a robotic system. Section 4 provides our experimentation. Finally, section 5 presents conclusions and some outlooks.

2. Attentive robots

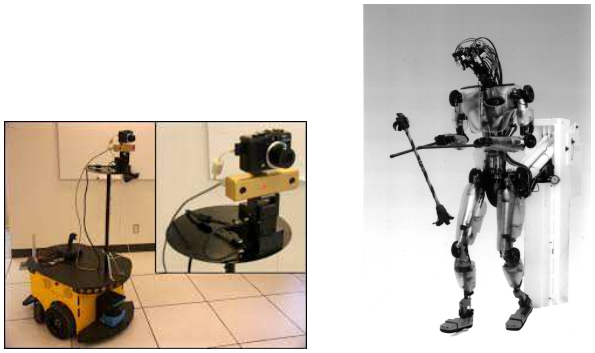
The tasks of the robot which involve visual attention might be classified roughly into three categories [5] further developed.

- low-level category: uses attention to detect salient landmarks that can be used for localization and scene recognition,
- mid-level category: considers attention as a front-end for object recognition,
- highest-level category: attention is used in a human-like way to guide the actions of an autonomous system like a robot.

In the first category robots use landmarks to compute their position in space. In [14] or [20], authors used static maps in which specific landmarks are located. In [6], the robot has to build a map and localize itself inside it at the same time. Salient regions are tracked over several frames to obtain a 3D position of the landmarks, and matched to database entries of all previously seen landmarks.

In the second category, attention methods are of special interest for all tasks in object detection and localization, or in classification of non pre-segmented images. [8] has recently integrated attentive object detection on the robot. In the same way, *Curious George*, developed by the laboratory for computational intelligence (Figure 2 (a)), was placed first in the robot league of the Semantic Robot Vision Challenge both in 2007 and 2008, and first in the software league for 2009[1] [12].

Finally, the highest-level category is dedicated to robots which have to act in a complex world facing the same problems as a human. One of the first active vision systems which integrated visual attention was presented by [3]. They describe how a robot can fixate and track the most salient regions in artificial scenes composed of geometric shapes. In [22], authors present an attention system which guide the gaze of a humanoid robot. The authors consider only one feature, visual flow, which enables the system to attend to moving objects (Figure 2 (b)). In [18], the humanoid robot *iCub* bases its decisions to move eyes and neck on visual and acoustic saliency maps. Others works concerning joint attention were done by [9] and [19].



(a) Curious George (b) 30DOF Humanoid Robot

Figure 2. Prototypes of attentive robots

As mentioned before many methods exist, but most of them need either strong information concerning the locations of landmarks or concerning objects to recognize. What we propose is to enhance a very tunable model which works in real time in order to integrate 3D information.

3. Our model and its extension

In this part we present the model we have developed. We classically used Laurent Itti's work [10]. The first part of its architecture relies on the extraction of three conspicuity maps based on low level characteristics computation that correspond to the production of information on the retina. These three conspicuity maps are representative of the three main human perceptual channels: color, intensity, and orientation. The second part of Itti's architecture proposes a medium level system which allows merging conspicuity maps (C^n), and then simulates a visual attention path on the observed

[1] http://google-opensource.blogspot.fr/2010/01/2009-semantic-robot-vision-challenge.html

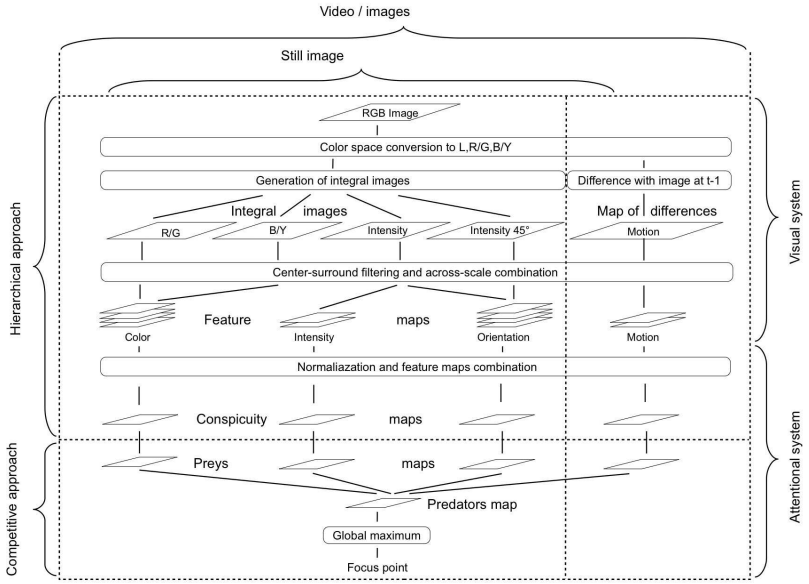


Figure 3. Our model of visual attention

scene. The focus is determined by a "*winner-takes-all*" and an "inhibition of return" algorithms. We have substituted this second part by our optimal competitive dynamics evolution equation [16], in which predator density map I represents the level of interest image contains and C^n represent respectively color, intensity and orientation prey populations *i.e.* the sources of interest, see Figure 3. For each of the conspicuity maps (color, intensity, orientation), the preys population C^n evolution is governed by the following equation:

$$\frac{d_{C^n_{x,y}}}{dt} = C^n_{x,y} + f\,\triangle_{C^{*n}_{x,y}} - m_C C^n_{x,y} - sC^n_{x,y} I_{x,y} \qquad (1)$$

with $C^{*n}_{x,y} = C^n_{x,y} + w{C^n_{x,y}}^2$ and $n \in \{c, i, o, m\}$, which means that this equation is valid for C^c, C^i, C^o andC^m which respectively represent color, intensity and orientation populations. w is a positive controlled feedback. This feedback models the fact that provided that there are unlimited resources the more numerous a population, the better it is able to grow. m^n_C is a mortality rate that allows to decrease the level of interest of regions in conspicuity map C^n. The population of predators I, which consume the three kinds of preys, is governed by the following equation:

$$\frac{d_{I_{x,y}}}{dt} = s(P_{x,y} + wI^2_{x,y}) + sf\,\triangle_{P_{x,y}+wI^2_{x,y}} - m_I I_{x,y} \qquad (2)$$

with $P_{x,y} = \sum_{n\in\{c,i,o,m\}} (C^n_{x,y}) I_{x,y}$.
This yields to the following set of equations, modelling the evolution of prey and predator populations on a two dimensional map:

$$\begin{cases} \frac{d_{C^i_{x,y}}}{dt} & = bC^i_{x,y} + f\,\triangle_{C^i_{x,y}} - m_{C^i} C^i_{x,y} - sC^i_{x,y} I_{x,y} \\ \frac{d_{I_{x,y}}}{dt} & = sC^i_{x,y} I_{x,y} + sf\,\triangle_{P_{x,y}} - m_I I_{x,y} \end{cases} \qquad (3)$$

As already mentioned, the positive feedback factor w enforces the system dynamics and facilitates the emergence of chaotic behaviors by speeding up saturation in some areas of the maps. Lastly, the maximum of the interest map I at time t is the location of the focus of attention. **This system has been implemented in real time**, see [4, 16, 17].

3.1. Extention to robotic environment

In order to enhance our model, and make it usable for robotic application, we have integrated with the previous model two new conspicuity maps. One for the depth and one for the isolated object.

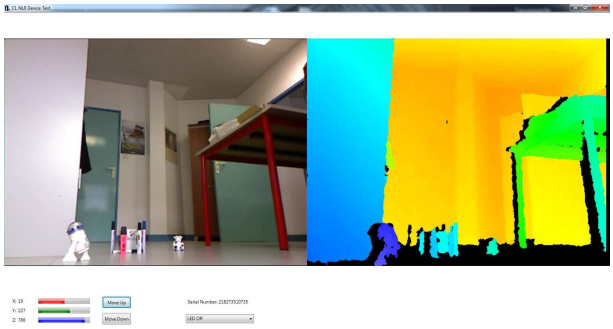
The depth conspicuity map. This map represents the depth of the scene in front of the robot. We have used a Kinect system from Microsoft which is a motion sensing input device. The SDK provides Kinect capabilities to developers to build applications which includes access to low-level streams from the depth sensor, the color camera sensor, and four-element microphone array. The depth sensor consists of an infra red laser projector combined with a monochrome CMOS sensor, which captures video data in 3D under any ambient light conditions. Let I_d be the depth image. Each pixel represents approximately the distance between Kinect and each object of the scene. In order to promote close objects rather than a distant one we define the depth consipicuity map as the inverse

of I_d.

$$C_d(i,j) = \frac{dyn_{C_d}}{dyn_{I_d} * I_d(i,j)} + \alpha \quad (4)$$

, where dyn_X represents the dynamic of image X and α a coefficient to constraint C_d to be positive. In order to avoid problems due to uncomputed depth in I_d, each null value on I_d stay null on C_d.

The isolation conspicuity map. This map has to focus on an isolated element. An isolated element is characterized by a pixel value different from its surroundings (lower or higher). In order to be as coherent as possible we have decided to use the same approach as the one used to detect information in the intensity conspicuity map. The only difference is that the input is not the intensity information but the depth map provided by the Kinect. Thus, we compute centre-surround differences to determine contrast, by taking the difference between a fine (center) and a coarse scale (surround) for the depth feature. This operation across spatial scales is done by interpolation to the fine scale and then point-by-point subtraction (Figure 4).



Original image and its depth map

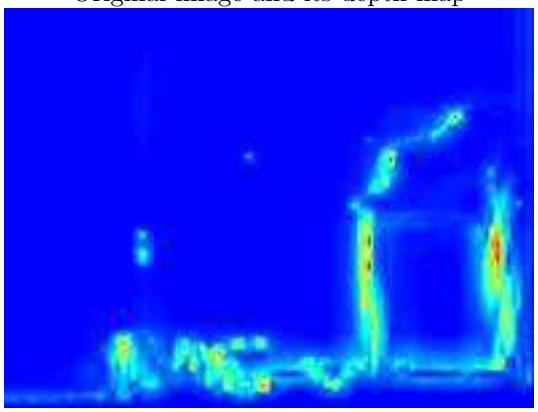


Figure 4. Isolation conspicuity map.

For each of the conspicuity maps (color, intensity orientation, depth and isolation), the prey population C^n evolution is governed by the following equation:

$$\frac{d_{C^n_{x,y}}}{dt} = C^n_{x,y} + f \triangle_{C^{*n}_{x,y}} - m_C C^n_{x,y} - s C^n_{x,y} I_{x,y} \quad (5)$$

4. Experimentation

For our experimentation we have decided to use a mobile system composed by a Lego Mindstorm system and a Kinect from Microsoft (Figure 5). The Lego Mindstorm allows motion, whereas Kinect allows video and depth acquisition. The Lego Mindstorms series of

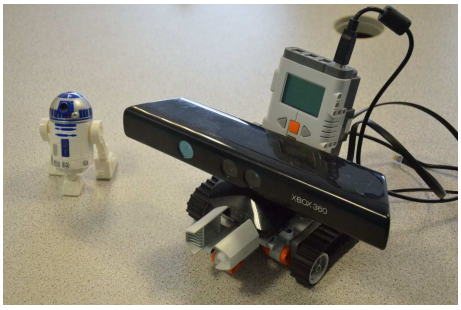


Figure 5. Our system composed by a Lego Mindstorm vehicle and a Kinect

kits contain software and hardware to create small, customizable and programmable robots. They include a programmable brick computer that controls the system, a set of modular sensors and motors, and Lego parts from the Technics line to create the mechanical systems. For our test we have decided to link the Lego Mindstorm and the Kinect to a computer thanks to a USB liaison rather than Bluetooth. Free tools in combination with the Robotics Developer Studio developed by Microsoft enable programming the Mindstorm using the C# language. Concerning Kinect, in 2011 Microsoft announced a non-commercial SDK to build applications with C#, see Figure 6.

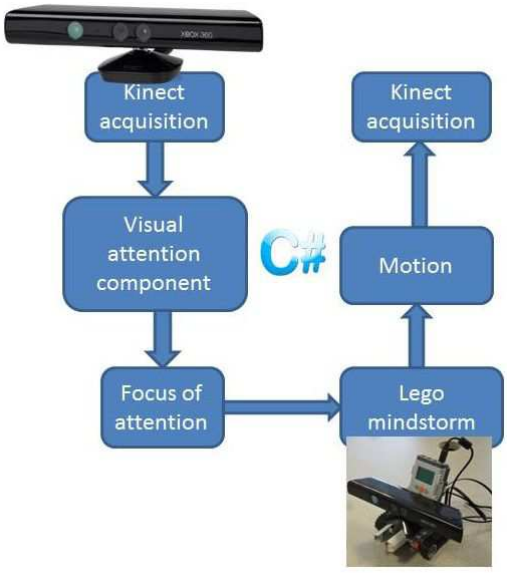
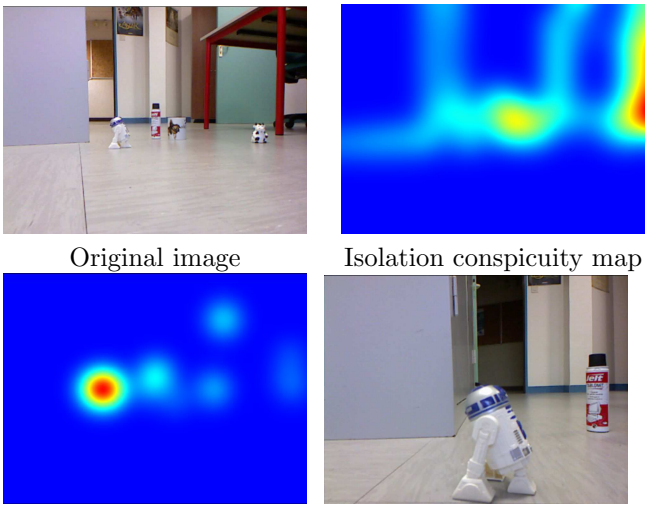


Figure 6. Block diagram of our system

Thus, we have decided to use C# to manage our application. It runs in real time on a computer DELL precision M4700 core i5 CPU 2.8 GHz, 8Go of RAM.
It is very difficult to evaluate our system. In fact we should evaluate the relevance of our results by using a headmounted eye tracking solution, and ask people to perform a very precise task. That's why we prefer assigning a specific objective to our system, and then subjectively evaluating the result. The objective assigned to our robot is to go to the *nearest salient* object and take a picture of it. An example is given Figure 7. We have done some experimentation in our lab, our office and hall. Figure 8 represents a small part of experiments we have realised and shows the relevance of our approach.



Original image | Isolation conspicuity map
Final saliency map | Picture taken by our robot

Figure 7. Presentation of different elements of our system.

5. Conclusion

This article proposes a robotic system which implements an attentive behaviour. This is a difficult task that has been addressed by only a few previous works, but represent an important milestone for the future. Attention is guided by a real time computational system inspired by [16] and modified in order to take into account depth and isolation. Our system is implemented thanks to a Lego Mindstorm robot and a Kinect. We have conducted very promising experiments and we would like to implement our system inside a Nao (Figure 9), an autonomous, programmable humanoid robot developed by Aldebaran Robotics. The first perspective we want to realize is to conduct a more global evaluation thanks to the NUS3D-Saliency Dataset provided by Tam V. Nguyen [11]. Moreover, we would like to integrate motion, in order to be reactive when a new element comes inside the robot field of vision.

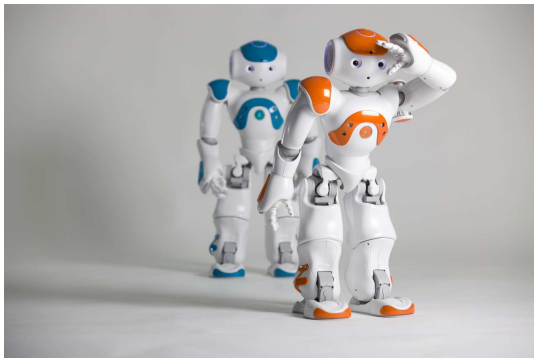


Figure 9. Nao looks at its futures capabilities.

References

[1] Allport, D. A. (1987). *Selection for action: Some behavioral and neurophysiological considerations of attention and action*, pages 395–419. Lawrence Erlbaum Associates, Hillsdale, NJ.

[2] Broadbent, D. E. (1958). *Perception and communication*. Pergamon Press, Elmsford, NY, US.

[3] Clark, J. and Ferrier, N. J. (1988). Modal control of an attentive vision system. In , *Second International Conference on Computer Vision*, pages 514–523.

[4] Courboulay, V. and Perreira Da Silva, M. (2012). Real-time computational attention model for dynamic scenes analysis: from implementation to evaluation. In SPIE, editor, *SPIE Optics, Photonics and Digital Technologies for Multimedia Applications*, page to be published, Brussels, Belgique.

[5] Frintrop, S. (2011). Towards attentive robots. *Paladyn*, 2(2):64–70.

[6] Frintrop, S. and Jensfelt, P. (2008). Attentional landmarks and active gaze control for visual SLAM. *IEEE Transactions on Robotics*, 24(5):1054–1065.

[7] Frintrop, S., Klodt, M., and Rome, E. (2007). A real-time visual attention system using integral images. In *5th International Conference on Computer Vision Systems (ICVS)*, Bielefeld, Germany. Applied Computer Science Group.

[8] Gould, S., Arfvidsson, J., Kaehler, A., Sapp, B., Messner, M., Bradski, G., Baumstarck, P., Chung, S., and Ng, A. Y. (2007). Peripheral-foveal vision for real-time object recognition and tracking in video. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, page 21152121, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[9] Heidemann, G., Rae, R., Bekel, H., Bax, I., and Ritter, H. (2003). Integrating context-free and context-dependent attentional mechanisms for gestural object reference. In Crowley, J. L., Piater, J. H., Vincze, M., and Paletta, L., editors, *Computer Vision Systems*, number 2626 in Lecture Notes in Computer Science, pages 22–33. Springer Berlin Heidelberg.

[10] Itti, L., Koch, C., Niebur, E., and Others (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

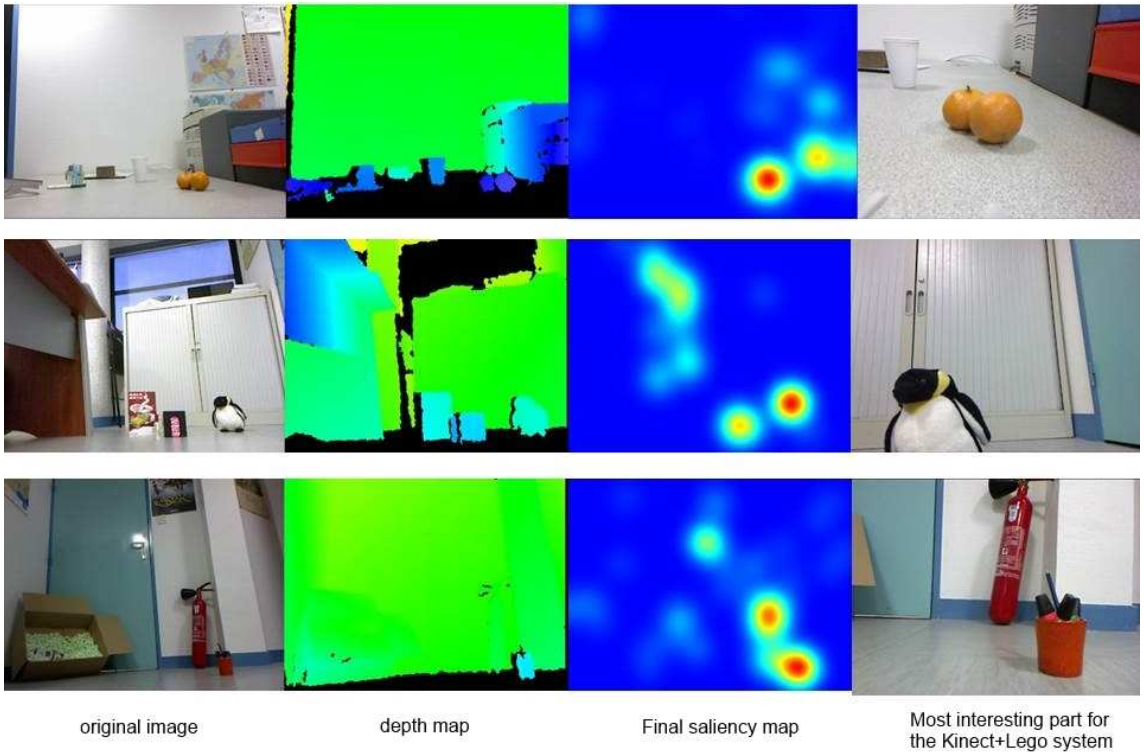


Figure 8. Sample of experiments realised

[11] Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., and Yan, S. (2012). Depth matters: Influence of depth cues on visual saliency. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *Computer Vision ECCV 2012*, Lecture Notes in Computer Science, pages 101–115. Springer Berlin Heidelberg.

[12] Meger, D., Forssén, P.-E., Lai, K., Helmer, S., McCann, S., Southey, T., Baumann, M., Little, J. J., Lowe, D. G., and Dow, B. (2007). Curious george: An attentive semantic robot. In *IROS 2007 Workshop: From sensors to human spatial concepts*, San Diego, CA, USA. IEEE.

[13] Neumann, O. (1987). *Beyond capacity: A functional view of attention. Perspectives on perception and action.*, pages 361–394. Lawrence Erlbaum Associates, Hillsdale, NJ, England.

[14] Nickerson, S. B., Jasiobedzki, P., Wilkes, D., Jenkin, M., Milios, E., Tsotsos, J., Jepson, A., and Bains, O. N. (1998). The ARK project: Autonomous mobile robots for known industrial environments. *Robotics and Autonomous Systems*, 25:83104.

[15] Ouerhani, N. (2003). *Visual Attention : From Bio-Inspired Modeling to Real-Time Implementation*. Thèse de doctorat, Université de Neuchâtel.

[16] Perreira Da Silva, M. and Courboulay, V. (2012). Implementation and evaluation of a computational model of attention for computer vision. In *Developing and Applying Biologically-Inspired Vision Systems: Interdisciplinary Concepts*, pages 273–306. Hershey, Pennsylvania: IGI Global.

[17] Perreira Da Silva, M., Courboulay, V., Prigent, A., and Estraillier, P. (2010). Evaluation of preys / predators systems for visual attention simulation. In *VISAPP 2010 - International Conference on Computer Vision Theory and Applications*, pages 275–282, Angers. INSTICC.

[18] Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., and Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *IEEE International Conference on Robotics and Automation, 2008. ICRA 2008*, pages 962–967.

[19] Schauerte, B., Richarz, J., and Fink, G. (2010). Saliency-based identification and recognition of pointed-at objects. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4638–4643.

[20] Siagian, C. and Itti, L. (2009). Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics*, 25(4):861–873.

[21] van der Heijden, A. H. C. and Bem, S. (1997). Successive approximations to an adequate model of attention. *Consciousness and cognition*, 6(2-3):413–28.

[22] Vijayakumar, S., Conradt, J., Shibata, T., and Schaal, S. (2001). Overt visual attention for a humanoid robot. In *Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on*, volume 4, pages 2332–2337 vol.4.