2014-10

Attention based object recognition applied to a humanoid robot

Joint Conference on Robotics and Intelligent Systems; Brazilian Robotics Symposium, 2th; Latin American Robotics Symposium, 11th; Workshop on Applied Robotics and Automation, 6th, 2014, São Carlos.
http://www.producao.usp.br/handle/BDPI/48623

Attention Based Object Recognition applied to a Humanoid Robot

Adam H. M. Pinto, Lucas O. de Oliveira,
Renata C. G. Meneghetti, Roseli A. F. Romero
Department of Computation
University of São Paulo
São Carlos, São Paulo, Brazil
E-mail: {adamh.moreira, lucas.orlandi}@gmail.com
{rcgm, rafrance}@icmc.usp.br

Alcides X. Benicasa
Department of Information Systems
Federal University of Sergipe
Itabaiana, Sergipe, Brazil
E-mail: alcides@ufs.br

***Abstract*—Analysis and recognition of objects in complex scenes is a demanding task for a computer. There is a selection mechanism, named visual attention, that optimizes the visual system, in which only the important parts of the scene are considered at a time. In this work, an object-based visual attention model with both bottom-up and top-down modulation is applied to the humanoid robot NAO to allow a new attention procedure to the robot. This means that the robot, by using its cameras, can recognize geometric figures even with the competition for the attention of all the objects in the image in real time. The proposed method is validated through some tests with 13 to 14 year old kids interacting with the robot NAO that provides some tips (such as the perimeter and area calculation formulas) and recognizes the figure showed by these children. The results are very promissor and show that the proposed approach can contribute for inserting robotics in the educacional context.**

***Keywords*-: Computer Vision, Robotics, Object-Based Recognition**

I. Introduction

Although our brain perceives a complex scene efficiently this is a quite demanding task for a computer. Evolution has developed ways to optimize our visual system in such a manner that only important parts of the scene undergo scrutiny at a given time. This selection mechanism is named visual attention [10], [17], [29].

According to [17], [30], several theories have been proposed and can be gathered in three main lines: location-based attention, feature-based attention and object-based attention. In this work, we consider that the selection visual is performed on object level, it means that the objects are the basic unit of perception. In this case, since the attention is directed to any part of an object, other parties also benefit from this attentional process [17], [30], [16].

Recent work has been conducted regarding to the knowledge of the target to influence the computation of the most salient region [28], [15], [5], [12], [14], [11], [7], [6], [3], [2], [4]. This knowledge is usually learned in a preceding training phase. The object-based visual attention model proposed in [4] has been extended and applied to the humanoid robot NAO to allow a new attention procedure to the robot, aiming to turn it able to recognize different planar geometric figures. In this paper, we measure this capability, with a mathematical question game, performed with a 13-14 years old kids. We hope that proposed approach contribute for inserting of robotics into educacional context.

This paper is organized as it follows. In section II, some works relating to this research are presented. In Section III, the description of computational method for attention based object recognition and the robot adopted in this work are explained. In Section IV, the experiments made to evaluate the proposed method are shown. In Section V, the results are discussed. Finally, Section VI, a conclusion and future works are presented.

II. Related Works

In this section, some works related to the use of artificial neural networks to recognise colors and objects are presented as well as some examples of humanoid robots interacting with children with autism spectrum disorder.

Robots are evolving from stationary devices that perform manufacturing tasks to mobile, information gathering, computing, and decision making platforms. In order to build autonomous robots that can carry out useful work in unstructured environments new approaches have been developed to building intelligent systems. Image recognition systems can be useful for a variety of automated-tasks, and, therefore, command considerable interest. A fast and highly robust vision system is very important in real-time object recognition [8].

In [20], [21] and [22], a MultiLayer Perceptron (MLP) artificial neural network with backpropagation algorithm was used to recognize colors in robot soccer domain. In a partially controlled environment (global vision obtained by ceiling cameras) the MLP in a RGB (Red, blue green) color space could recognize all 7 colors needed in soccer game. However, in different brightness conditions the HSL (hue, saturation and lightness) color space could obtain more accuracy and preciseness and decrease the execution-time of recognizing an object from images [24].

Waldherr [26] developed an interface to gesture recognition for controlling a mobile robot with a manipulator. The system uses a camera to track a person and recognize arm motion, allowing the robot follows reliably a person with changing lighting conditions. In [25], an accuracy of 98.5% was obtained in recognizing gestures to control an mobile robot.

978-1-4799-6711-7/14 $31.00 © 2014 IEEE
DOI 10.1109/SBR.LARS.Robocontrol.2014.19

An experiment with four autistic kids was proposed in [23], comparing the human-robot interaction and human-human interactions in a motor imitation task. In real-time, the robot NAO imitated gross arm movements, and different behavorial criteria were analysed: eye gaze, smile/laughter, gaze shifting. While two children did not mind with the robot's presence, the other two showed more smile and eye gaze, compared to the human partner.

In [13], it was tested the communication and social skills in a memory card face matching game. Adolescents with autism and with other cognitive impairments are recruited in pairs (one of each pair had autism) and the game was played in three different game modes (using robot, smart boards and playing cards) for aproximately 15 minutes in three separeted days. Repetitive behaviors was reduced in participants with autism when using both robot and smart board. It shown that it is feasible to use a robot to assist teaching of social skills to adolescents with autism, but suggest that the robot features could be further explored and utilized.

Another experiment to test the potencial application of humanoids robots was performed at the Children's Hematology and Stem Cell Transplantation Unit of Szent László Hospital, in Budapest [9]. Forced to live in a 2x3 m sterile boxes, the robot NAO was a good companion to cheer kids up and to do some exercises. This is a new and promising domain of cognitive infocommunications.

As the robots are accepted by children in general, we intend to verify if the humanoid NAO would be also accepted into educacional context.

III. Methodology

In this section, are presented all of methods needed to turn the robot NAO capable to realize visual attention based on object recognition.

A. Top-down Biasing and Modulation for Object-Based Visual Attention

The visual attention model proposed in [4] is composed by the following modules: a visual feature extraction module, a top-down biasing feature-based, a LEGION network for image segmentation, a network-based high level data classifier for object recognition, a network of integrate and fire neurons, which creates the object-saliency map and, finally, an object selection module, which highlights the most salient objects in the scene.

The first stage in visual attention model is responsible for extracting the early visual features in parallel across the scene. The results from this stage are the following conspicuity maps: colors, intensity and orientation. In this work, we consider only the color channel. The next stage of the model is the combination of the results from the conspicuity maps with specific weights, for the top-down biasing of the LEGION segmentation network. The implementation of the LEGION followed the algorithm proposed in [27]. The output from those modules feed the following modules: the network for object recognition and the network for integrating and firing neurons, which creates the object-saliency map.

The top-down biasing is defined by the association of weight to output from the conspicuity map (C_c). The saliency value for conspicuity map is weighted and combined into a saliency map S_m defined as:

$$S_m = \frac{1}{n^c} W_c C_c, \tag{1}$$

where n^c denotes the conspicuity map and W_c determines weight of the conspicuity map C_c.

According to [27], the segmentation process in the LEGION is based on the idea that a segment must contain at least one oscillator, denoted as a *leader*, which lies in the center of a large homogeneous region. Leaders are all oscillators i in which the lateral potential $p_i \geq \theta$ where θ is a threshold [27]. In order to generate the top-down biasing of the proposed model, an oscillator i defined as leader only will pulse if its saliency value $S_{m_i} \geq \theta_{bias}$.

The proposed model takes both bottom-up and top-down modulations into account. Early visual features, i.e. color contrast, define the bottom-up signal. On the other hand, information about previously memorized objects and their features (top-down modulation) is responsible for guiding the selection process. Thus, in order to apply the proposed model to select the salient objects of a given scene, the MLP network must be trained with a set of objects representing the desired targets of the scene.

(a) (b) (c) (d)

Figure 2. Samples of objects for training the object recognition module.

After the training process, MLP network is able to recognize a set of segments (objects). Thus, the overall dynamics of the system can be understood as it follows (see Figure 1). Each time a segment is highlighted (pulsing) into LEGION network, it is directly presented to MLP network. The output of the MLP indicates whether or not the object is among those memorized by the recognition system. If the object is recognized, the network output value is used for setting the attribute recognition parameter $R_{i,j}$, where i and j represent the spatial position of pixels inside each segment. Initially, $R_{i,j} = 0$ for all neurons. At the end of this process, all the neurons related to the objects, that should receive attention (*top-down* modulation), will be assigned to a recognition value ($R_{i,j} = [0,1]$), that will modulate the attentional process. Segments representing unknown objects can also present nonzero recognition values. In order to avoid those objects receiving top-down modulation, a threshold for the recognition value (θ_r) is adopted. Thus, segments below this threshold are not considered. Hence, the value of recognition $R_{i,j}$ is defined by:

$$R_{i,j} = \begin{cases} 1, & \text{if } R_{i,j} \geq \theta_r \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

The proposed method can recognize even a nonlinearly separable objects in a scene. In Figure 3, the method

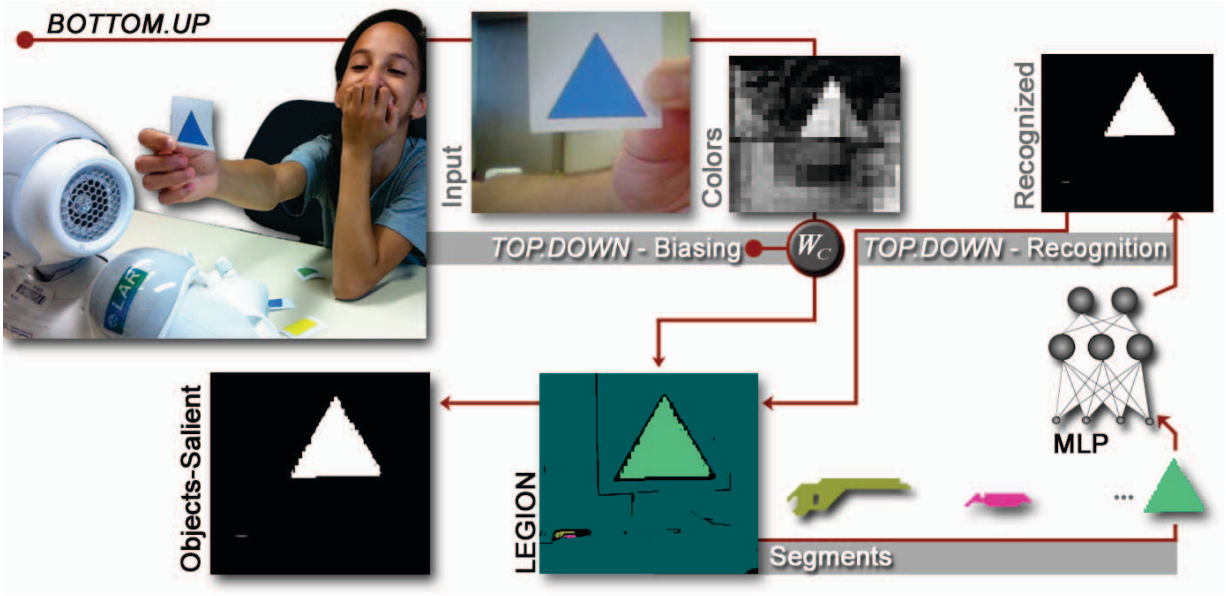

Figure 1. A flowchart of object-based selection.

correctly could find two spirals, what is not a trivial accomplish. On each pulse of the network, the method will select a salient segment in the scene and can even detect overlapped figures, unlike the system that is already available into NAO robot simulator. Also, the visual system can recognize figures in real time and in dynamical environments. Although, in this paper, the application will not require that nolinearly separable objects or overlapped figures, we decided to adopted the presented approach because we intend to elaborate other more complex applications directed by students as a future work. To be more precise, we are addressing the ability of the robot NAO to recognize plane geometric figures and the human-robot interaction feedback.

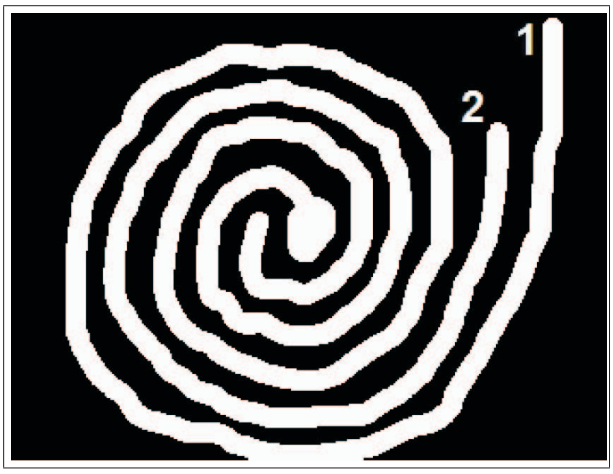

Figure 3. Nonlinearly separable figure

B. Humanoid Robot NAO

NAO, from Aldebaran Robotics, is a 58-cm tall humanoid robot that can move, recognize face and speech, and can talk. With a visual programing language, NAO can be used in computer and science classes even with primary school children. Considering the state-of-art research in [19], that shows good results in human-robot interaction with autistic children, we decided to use the robot to evaluate the vision algorithm presented in the previous section, in a real world, and its interaction skills.

NAO comes with a embedded software running in the head of the robot, allowing autonomous behaviors. It is called OpenNAO, a GNU/Linux distribution based on Gentoo, specifically developed to the robot. OpenNAO provides numbers of libraries and programs, but the main software is the NAOqi, the software that allows the robot to move, an so on.

C. NAOqi

Some desktop softwares allows the creation of new behaviors and the remote control of the robot, running on a computer. There is the monitor, but is only dedicated to give some feedback, and the Choreograph, a visual programming language. With Choreograph is possible to create and test animations and behaviors, trying it on a simulated robot before use the real one. Those behaviors are written in a graphical language and NAOqi interprets and execute them. We used Choreograph to provides all the interaction with the kids during the experiments, as the moviments, voice and coloring of the robot's LED.

D. C++ SDK

To use our own software, we needed to use one of the available SDK, like is presented in Figure 4. A new module was created and uploaded on the robot, with the

algorithm presented. The NAOqi API is currently available in 8 languages, but only C++ and Python are supported on the robot. C++ is the most complete framework and is the only one that let's write a real-time code, what is essential in our approach.

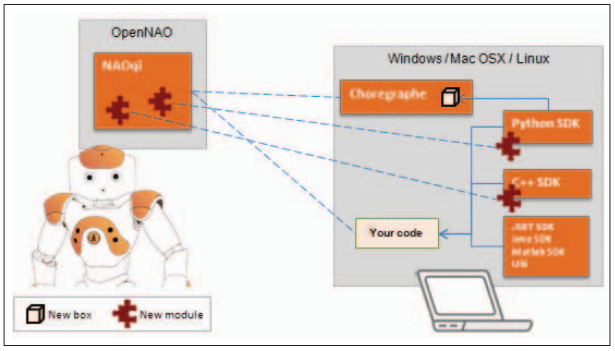

Figure 4. How Nao Software Works [18]

To compile this code is recommended to use CMake with qiBuild framework. qiBuild manages dependencies between projects and supports cross-compilation. It is mandatory to compile exclusively for 32-bits architectures, but the result code will run as well on 32-bits or 64-bits bit environment. Using NAOqi C++ SDK supports OpenCV 2.3.1 (in our robot version) and further we used a QTCreator IDE to implements the attention based vision to robot NAO.

IV. Experiments

In the experiments for testing the entire system, we decided to make a geometric figure question game with children. We separated the children in two groups. The first group would do the test first interacting with robot NAO and after with humans. The other group would do the opposite. It is always three different figures and three tips to each one were presented for the students. In figures 5 and 6, one can see these experiments.

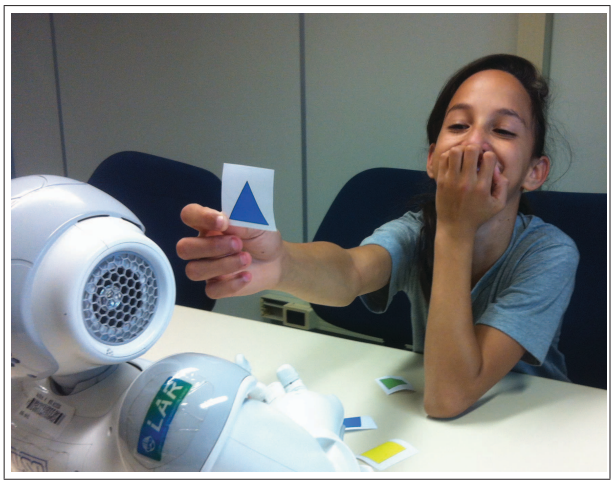

Figure 5. Experiment with kids

In the experiments, we use a Mean Opinion Score (MOS) [1] technique to do a subjective evaluation. The technique offers a scale to measure the quality of the interactions with NAO, its movements and the ability to understand and give a correct answer according to the geometrical figure shown. This scale ranges in:

- 1 - Bad
- 2 - Poor
- 3 - Fair
- 4 - Good
- 5 - Excellent

A total of 22 volunteers, students ranging from 13 to 14 years of age, and two of which have low autism, play a question game. To begin, eleven kids interact with the robot NAO, a hint is given to each kid by the robot such as the internal angles of the geometric figure, and it was expected the child to show to the robot the correct object. The robot recognizes the figure shown and compare if is the expected figure. If it is correct, the robot blinks, and explains the figure to the kid. If it is wrong, NAO's eyes turn to red and it gives another tip (like the perimeter formula). This process is continued if the student answered wrong again, then there is one last chance. If the student answered incorrect once again, NAO will tell him what figure it was and then will restart the game with another figure. The speech and the reaction of the robot are predefined, but throughout the test, the robot is acting autonomously, recognizing the shown figure and defining by the appropriate reaction (if the answer given by the student is correct or not).

Three questions are considered after the game to define the MOS quality:

- What do you think about the timing of the robot's responses and actions?
- Did you understand everything that robot said?
- Did the robot answer wrong during the test?

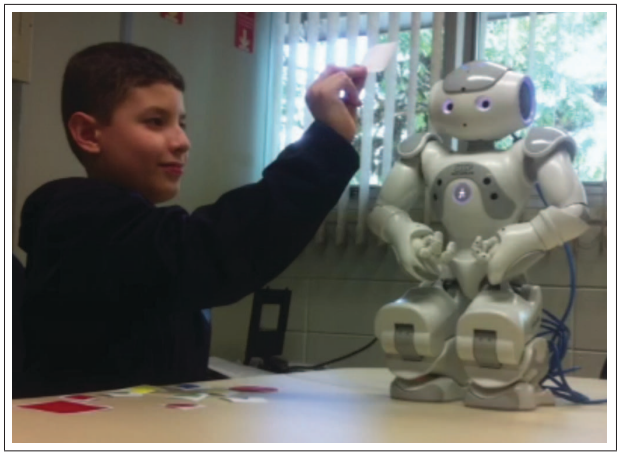

Figure 6. Experiment with kids 2

After that, those students were lead to another room, and did the same game without the robot. To other eleven children, we did the opposite. Firstly, the interaction

occurred without the robot and, in a second moment, with NAO. All the interaction was recorded in video and all the expressions and responses of those children were considered to the final result. Any observation about the experiment could be write by the volunteers. At the end some new questions are considered:

- Now that you finished this test, what do you think about robotics?
- Do you think that a robot could help you in your studies?
- What do you think about having a robot teacher?
- Do you wish to do another test with robots?

V. Discussion of Results

In total, NAO answered 52 times in all of the experiments. Only one student managed to hit all three figures in the first attempt, forcing the robot to give more tips for all others. The confusion matrix is presented in Table I. In all ot the tests, the robot had an precision of 75% and an accuracy of 78% for recognizing the figures shown.

Table I
Confusion Matrix of Robot Responses

Confusion Matrix	True	False
Positive	15	5
Negative	26	9

About the MOS quality questions, 36% of students rates 5 (Excellent) to the robot actions and responses, 27% rates 4 (good), 27% rates 3 (Fair) and 10% rates 1 (Bad). 77% rates 5 to robot's voice (they could understand everything that NAO said) and 23% rates 4.

Although 100% of children said they would participate in more experiments with robots, 27% of them had doubts about having a robot as a teacher. Among the problems mentioned, they pointed out a lack of investment for buying/building robots and the need to improve the robots performance.

NAO was not ready to repeat questions when they did not understand what it was talking, when it gave a false positive or false negative answer. The children realized the need to have someone helping when the robot gives some problem. It was also noted that illumination problems and noise hinder the recognition process. In the example of Figure 7, we had a poor segmentation result, based on the competition for attention with the figure and the hand.

All children gave 100% attention to the robot throughout the interaction, including the two with autism. For these two in particular, when the robot was off, they were interested in everything and they could not be quiet in their chairs. But, when the robot starts moving and talks, they just gave all attention to the robot. But one of them just tried to guess the figures, not caring about the tips. Always asking how the robot works, he seemed more interested in the robot than the math questions.

In relation to the robot performance, we have that the robot had a good performace as compared to the results

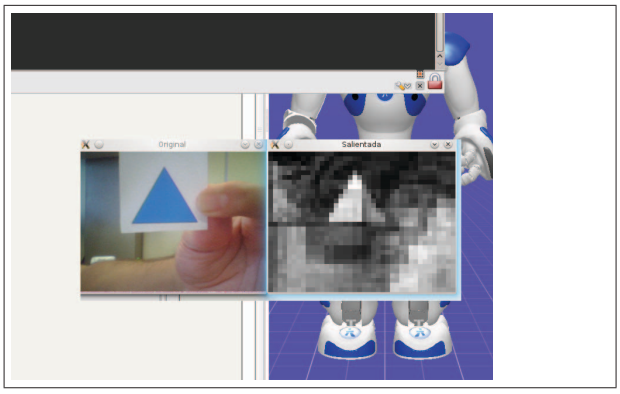

Figure 7. Poor NAO segmentation result

obtained in the lab tests, for plane figures, without to use the robot NAO, presented in [4].

VI. Conclusion

In this work, it was demonstrated that attention based object recognition method adopted was efficient to predict the salient objects in dynamical environments. Thanks to this vision system, the humanoid robot NAO could recognize figures and interact with students with 13 to 14 years old.

The MOS quality test showed that the robots actions need to be improved. Some children claimed that the robot was talking too fast or too low and they could not understand everything. Also, the robot was not prepared to give the same tip twice or restart the game when given a false negative or false positive response (the robot did not recognize the figure shown by the student).

An interesting results was obtained, the eleven students that interacted with the robot first, during the second test (with the human professor) they scored better than the other eleven students who did the opposite. They enjoyed to work with the robot and paid more attention than those who were with the teacher first. The robot NAO had a good accuracy in recognizing figures in dynamical environment. Therefore, the insertion of robots for helping in the learning process was very well received by the students. The robot NAO provided a more interaction between student and the study object

As future works, we intend do a mathematical game involving nonlinearly separable figures for evaluating the performance of the vision system and analyzing the interaction quality of the robot NAO with students.

Acknowledgment

This work was supported by the São Paulo State Research Foundation (FAPESP) and the Brazilian National Research Council (CNPq).

References

[1] *Methods for subjective determination of transmission quality. CCITT Recommendations; electronic version*. Recommendations. IUT, Geneva, 2000.

[2] Alcides X. Benicasa, Marcos G. Quiles, Liang Zhao, and Roseli A.F. Romero. An object-based visual selection model with bottom-up and top-down modulations. In *Neural Networks (SBRN), 2012 Brazilian Symposium on*, pages 238 –243, Curitiba,PR-Brasil, oct. 2012.

[3] Alcides X. Benicasa, Liang Zhao, and Roseli A.F. Romero. Model of top-down / bottom-up visual attention for location of salient objects in specific domains. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1582–1589, Brisbane-AU, june 2012.

[4] AlcidesXavier Benicasa, MarcosG. Quiles, Liang Zhao, and RoseliA.F. Romero. Top-down biasing and modulation for object-based visual attention. In Minho Lee, Akira Hirose, Zeng-Guang Hou, and RheeMan Kil, editors, *Neural Information Processing*, volume 8228 of *Lecture Notes in Computer Science*, pages 325–332. Springer Berlin Heidelberg, 2013.

[5] J. Bonaiuto and L. Itti. Using attention and spatial information for rapid facial recognition in video. *Image and Vision Computing*, 24(6):557–563, 2006.

[6] Ali Borji, Majid N Ahmadabadi, and Babak N Araabi. Cost-sensitive learning of top-down modulation for attentional control. *Machine Vision and Applications*, 22(1):61–76, 2011.

[7] Ali Borji, Majid Nili Ahmadabadi, Babak Nadjar Araabi, and Mandana Hamidi. Online learning of task-driven object-based visual attention control. *Image and Vision Computing*, 28(7):1130–1145, 2010.

[8] A. K. Chakraborty, D. Pal, and P. Chatterjee. Fast recognition of mechanical objects using neural networks under robust aspect. In *Journal of The Institution of Engineers (India) vol. 93(1)*, pages 55 – 62, 2012.

[9] E. Csala, G. Nemeth, and Cs. Zainko. Application of the nao humanoid robot in the treatment of marrow-transplanted children. In *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, pages 655–659, Dec 2012.

[10] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

[11] L. Elazary and L. Itti. A bayesian model for efficient visual search and recognition. *Vision Research*, 50(14):1338–1352, Jun 2010.

[12] Simone Frintrop. *VOCUS - A Visual Attention System of Object Detection and Goal-directed Search*. PhD thesis, PhD thesis, Lecture Notes in Artificial Intelligence (LNAI), 2006.

[13] Kimberlee Jordan, Marcus King, Sophia Hellersteth, Anna Wiren, and Hilda Mulligan. Feasibility of using a humanoid robot for enhancing attention and social skills in adolescents with autism spectrum disorder. *International Journal of Rehabilitation Research*, 36(3):221–227, 2013.

[14] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimal object detection.

[15] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005.

[16] Kathleen M. O'Craven, Paul E. Downing, and Nancy Kanwisher. fmri evidence for objects as the units of attentional selection. *Nature*, 401:584–587, 2005.

[17] H. Pashler. Introduction. in h. pashler (org.). *Attention. Hove (Reino Unido): Psychology Press.*, 1998.

[18] Aldebaran Robotics. Nao software 1.14.5 documentation. 2014.

[19] S. Shamsuddin, H. Yussof, L. Ismail, F.A. Hanapiah, S. Mohamed, H.A. Piah, and N. Ismarrubie Zahari.

[20] A. S. Simoes and A. H. R. Costa. Segmentação de imagens por classificação de cores: Uma abordagem neural para representaçcão rgb. In *C. H. C. R. M. T. S. Sakude (Ed.), Workshop de computacao WORKCOMP, ITA-Sao Jose dos Campos*, pages 25 – 31, 2000.

[21] A. S. Simoes and A. H. R. Costa. Using neural color classification in robotic soccer domain. In *Internacional Joint Conference IBERAMIA*, 2000.

[22] A. S. Simoes and A. H. R. Costa. Classificação de cores por redes neurais artificiais:um estudo do uso de diferentes sistemas de representação de cores no futebol de robôs móveis autônomos. In *ENIA*, 2001.

[23] A. Tapus, A. Peca, A. Aly, C. Pop, L. Jisa, S. Pintea, A.S. Rusu, and D.O. David. Childrens with autism social engagement in interaction with nao, an imitative robot: A series of single case experiments. *John Benjamins Publishing Company*, 13(3):315–347, 2012.

[24] S. Tsai and Y. Tseng. A novel color detection method based on hsl color space for robotic soccer competition. In *Computers and Mathematics with Applications*, pages 1291 – 1300, 2012.

[25] C. Tzafestas, N. Mitsou, N. Georgakarakos, O. Diamanti, P. Maragos, S. E. Fotinea, and E. Efthimiou. Gestural teleoperation of a mobile robot based on visual recognition of sing language static handshapes. In *IEEE International Symposium on Robot and Human Interactive Communication*, pages 1073 – 1079, 2009.

[26] S. Waldherr, S. Thrun, and R. A. F. Romero. A gesture-based interface for human robot interaction. In *Autonomous Robots 9(2)*, pages 151 – 173, 2000.

[27] DeLiang Wang and David Terman. Image segmentation based on oscillatory correlation. *Neural Computation*, 9:805–836, 1997.

[28] Jeremy M. Wolfe and Todd S. Horowitz. What attributes guide the deployment of visual attention and how do they do it ? *Nature Review Neuroscience*, 5:495–501, 2004.

[29] Steven Yantis. *Attention*, chapter Control of visual attention, pages 223–256. Psychology Press, London, 1998.

[30] Steven Yantis. *Attention and Performance XVIII*, volume 18, chapter Goal-directed and stimulus-driven determinants of attentional control, pages 73–103. MIT Press Cambridge, 2000.