Active attention for human robot interaction using visual and depth information

F. Javier Page Alcalde, Ismael García-Varea and Jesús Martínez-Gómez

***Abstract*—This paper proposes an attention system for robot human interaction. This system uses depth and visual information, provided by a Kinect device, and has two main objectives: detect when a user is requiring attention and predict the objects of the environment that are suitable for interaction. The hypothesis of the article is that the attention system can provide robots with the necessary information to start a conversation with humans. The work details all the techniques that have been used to detect a human pose and some specific movements in order to start a communication with a robot, and to extract shape-based and colour-based features from the environment the human and the robot intetact with. Experiments on human pose estimation and object detection in real world are reported in order to demonstrate the feasibility of the proposed system.**

***Index Terms*—Computer vision, robotics, active vision, human robot interaction, cognitive robotics.**

I. Introduction

ROBots have traditionally had little interaction with humans. In fact, for a long time most robots were only found in industry and the only communication consisted in planning and supervising the work of the machine by a human operator. However, given the paradigm shift in research with robots and the rise of social robotics, it is necessary to rethink strategies for achieving different mechanisms of effective HRI [1]. Interaction, by definition, requires communication between two entities: humans and machines. Here we consider the interaction as a social and close interaction, i.e. an interaction in which the two entities are in the same physical space, taking into account social, emotional and cognitive aspects. This type of communication is therefore essentially multimodal, which is produced using stimuli of various kinds. As an example, humans communicate primarily using speech (physically speaking via an acoustic signal), but instinctively we add much more information such as body posture and limb movements (indicating direction, size, etc.), and even gestures or facial expressions, which often provide very relevant information in communication.

The literature quoted Isaac Asimov and his laws of robotics as the beginning of HRI. However, the technology boom began in the second half of the 90's and at the beginning of the new century. Since then, many research groups focused their attention on this emerging area. From the standpoint of a social robot, and since it has to communicate with people following behaviours, patterns and social norms, it needs to have skills that are within the domain of the so-called social intelligence [1]. We have to bear in mind that the socializing of a robot with people is a difficult issue, mainly because robots and humans do not share a common language and perceive the world the same way. In order to interact properly, social robots need to be able to communicate with people holding high-level dialogues. For this, they must meet certain requirements. First, the robot must be able to visually track the movements of the interlocutor [2]. In this context, the active sensing system has to be used in order to discriminate relevant information obtained by the robot from other information captured by the cameras. Also, the robot must be able to recognize and interpret human speech, including affective speech, discrete commands and natural language [3]. In addition, the social robot must have the ability to recognize facial expressions, gestures and human actions and interpret the social behaviour of people through the construction of elaborate cognitive-affective models [1]. In this work we are going to be centred on the first problem, leaving the other two problems for future developments.

In the literature we find works that incorporate multimodality in the process of communication between men and robots, in which they fuse different sources of knowledge such as speech and gestures of the user. In this regard it is worth noting the work proposed in [4], where the use of an ontology for a mobile robot to learn and remember areas and objects within a controlled environment is proposed. Similarly, in [5] a project with speech and gestures between a user and a group of robots is described, and in [6] a framework for spoken dialogue interfaces applied to the previous scenario is defined.

The hypothesis of this paper is to provide a robot with the initiative to start a conversation with a human, enhancing the naturalness of the HRI. For this purpose, the robot have to be able to detect when a user is requiring attention, and acting in anticipation asking about what object in the environment it has to manipulate and interact with. To support this hypothesis several experiments have been carried out. The obtained results show us that the approach proposed here is appropriate as a starting point for a more complete (including bidirectional speech communication) multimodal human robot interaction system.

The rest of the paper is structured as follows: in section II, to pave the ground, a brief review of multimodal HRI is presented. Section III is devoted to describe how the visual attention has been carried out. In section IV our active vision system for HRI is presented. In section V the experimental framework is described as well as the achieved results. Finally, in section VI the major conclusions and future research

F. Javier Page, Ismael Garcia-Varea and J. Martinez-Gomez are with the University of Castilla-La Mancha, Albacete, Spain
E-mail: fco.jpage@hotmail.com, Ismael.Garcia@uclm.es and Jesus.Martinez@uclm.es

directions are drawn.

II. Multimodal human-robot interaction

In the literature we can find multiple definitions of what Human-Robot Interaction (HRI) is or is intended to be. HRI, basically, is devoted to the understanding, design and evaluation of robotic systems for their use by or with humans[7]. Going a little bit further, Multimodal HRI is the study of interactions between humans and robots using multiple sources of information to provide a natural way of communication.

The interaction between robots and humans is influenced by the proximity to which they are. According to that, HRI can be classified as remote and proximate interaction. In one hand, remote interaction is considered when the human and the robot are separated temporally and spatially. This type of interaction focuses on teleoperation, supervisory control and remote manipulation (or telemanipulation). On the other hand, proximate interaction is considered when both human and robot are placed in the same scene, and is mainly focused on the so-called social robots, which includes social, emotional and cognitive aspects of interaction[7]. The types of communication that exist in this interaction can be grouped into: oral, visual and gestural.

Here we are focused in the proximate multimodal interaction approach, where a social mobile robot tries to communicate with a human, with the final goal of localize, recognize, and manipulate objects within the environment where the robot and the human are interacting simultaneously.

III. Visual Attention

Social robots are required to interact with people in complex environments like houses or hospitals. These robots should perform several tasks in real time and they have to select specific regions in the scene for further analysis. The selection of the region in the scene will depend on the purpose of the robot, but we identify two main objectives for selecting a specific region of the scene. The first one is to determine the person who requires more attention from the robot. The second objective consists on locating suitable interest points.

These two objectives could be achieved together to improve the human robot interaction in the following way. Firstly, the robot focuses the attention on the person who requires more attention. After that, it selects the interesting objects of the environments that could be used to interact with such person.

The detection of the interest of the humans in establishing an interaction with the robot is performed by processing several channels. Humans use sounds and gestures, along with their facial expressions, to denote interest in the interaction. All these sources of information have been used in several works as [8] or [9], while other systems are mainly based on visual information [10].

Concerning the problem of locating interest objects to interact with, most approaches like [11] or [12], rely on the theory of visual attention in humans [13]. These proposals attempt to predict which location in the images acquired by the robot would focus the attention of a human. In order to achieve this goal, the input image is processed to generate a saliency map. The generation of this map includes a decomposition of the input image using features based on colour, intensity and orientation. The complete process for detecting the focus of attention is shown in Fig. 1[1].

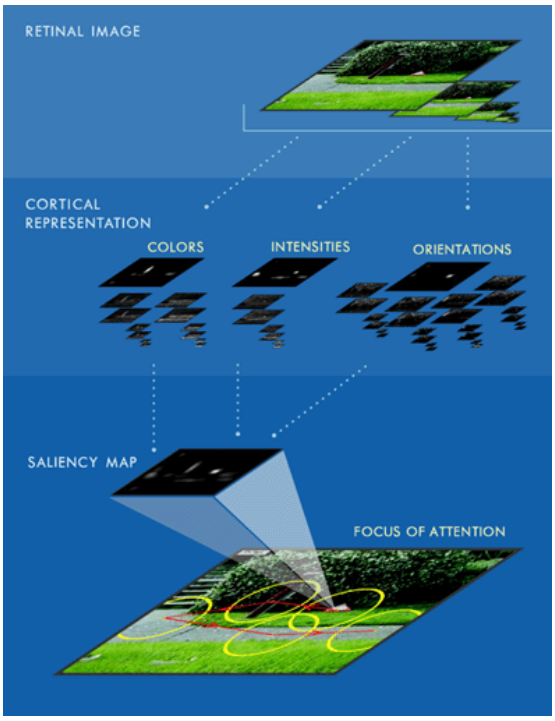

Fig. 1. Visual attention process based on colour, intensisties and orientation.

In this work, we decided to focus on locating suitable interest object in the environment, assuming that the robot is interacting with the person who requires more attention. We are not only interested on locating the most interesting area of the image but also on extracting information from this area. The information retrieved from the scene are the objects that are suitable for being used in the human robot interaction process. Thanks to this knowledge, the robot can play an active role in the communication with humans by proposing candidate objects for interaction.

Moreover, the robot can ask information to the human about the candidate objects it detects. The robot can learn from this dialog several characteristics from an object: its description, how can it be manipulated or how important is it for humans. All this information can provide the robot with the ability to recognize, interpret and represent the scene in a way comprehensive to humans [14].

IV. Active vision system for HRI

This section describes the complete vision system that has been developed. The objective of this system is to detect the objects of the environment suitable for being used to interact with a human agent. It uses as input the visual and depth images provided by a Microsoft Kinect device. The depth information has been used to detect the pose of the human the

[1] obtained from http://ilab.usc.edu/bu/theory/index.html

robot is interacting with. The visual information was processed to locate the objects of interest in the scene. Both sources of information are combined to perform better: the pose of the human allowed discarding useless information in the visual image. All this process is illustrated in Fig. 2.

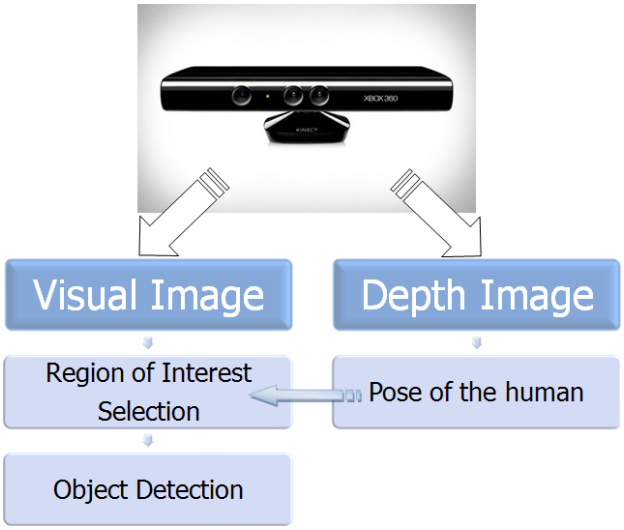

Fig. 2. Process of the whole attention system.

A. Depth Information

The systems processes the input depth image by using the OpenNI framework[2]. Concretely, we took advantage of the body tracker method provided by OpenNI, which detects in real time the position of the head, the arms, the body and the legs. The tracking process starts when a human located in front of the Kinect device has the PSI pose. The PSI pose and the positions of the human body that are detected with the Kinect device can be observed in Fig. 3. The PSI pose is also used in OpenNI to calibrate the biometrics of the human to track, improving the accuracy for the future detections.

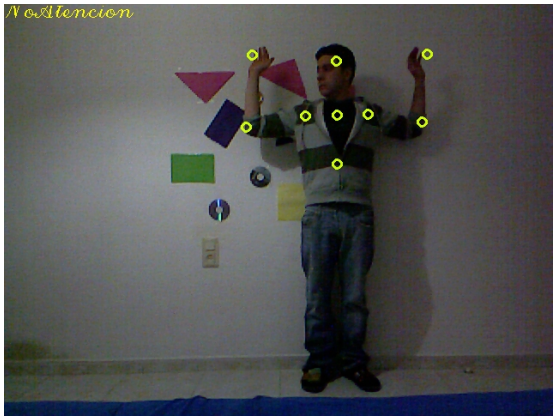

Fig. 3. PSI and body points detected with the Kinect device.

The data acquired with the Kinect device was used to detect the pose of the user. There are more locations of the human body that can be detected, but we decided to use only the 9 that are shown in Fig. 3. Three different poses for the human are identified in this work: "NoAttention", "AttentionLeft" and "AttentionRight". "NoAttention" denotes a user that is not requiring attention from the robot. "AttentionLeft/Right" is used to describe a user that is requiring to interact with an object located to the left/right with respect to the robot, and this pose is identified by the position of the head and the arms (pointing left or right respectively).

1) Classifier: In order to detect the pose of the user, we generated a classifier from the data acquired with the depth image. There are 28 different features that come from the 9 three-dimensional locations of the body ($< x, y, z >$ values) and the orientation of the head. We generated a balanced dataset with 477 samples by recording a human with the Kinect device. After generating the database, we performed a feature selection step using Weka [15] and the InfoGainAttributeEval attribute selection method. As a result of the feature selection, only 11 attributes were maintained.

The next step consisted on generating the classifier. We used a decision tree classifier (C4.5 or J48 in Weka) that was created using a 10-fold cross-validation training criteria, obtaining an accuracy rate of 85.9%. More details about the classifier are given in Section V

B. Visual Information

1) Region of interest: After detecting the pose of the human that interacts with the robot, the system processes the visual information. In order to reduce the amount of data to work with, we discarded all the useless pixels from the visual image acquired by the camera. This process was done by taking into account the pose and the location of the head of the human in the image.

The position of the head determines the starting point for the region of interest to extract. This region represents where the human is looking at and we decided to use a cone for representing this field of view. The vertex of the cone corresponds to the position of the head of the human while its orientation depends on the pose and the position of the right/left hand. It should be pointed out that the position of the head and the hands is provided by the depth camera and, therefore using three dimensional coordinates. This implies a projection of the 3D world points to the visual image coordinates system, which was done using Eq. (1) and (2). The value of the focal distance (f_x, f_y) and the optical centre (c_x, c_y) was set to (594,591) and (330,270) respectively.

$$x_c = \frac{x_w \cdot f_x}{z_w} + c_x \tag{1}$$

$$y_c = ((\frac{y_w \cdot f_y}{z_w}) - 1) + c_y \tag{2}$$

The cone is created using a 45 degrees angle with their sides oriented to the left/right hand of the human, depending on the pose. An example of a cone is shown in Fig. 4, where the pose "AttentionLeft" was detected . All the pixels of the image which are outside of the region (cone) of interest are discarded, which reduces the complexity of the visual processing algorithms.

[2]http://www.openni.org/

Fig. 4. Extraction of the region (cone) of interest.

2) Object detection: After extracting the region of interest, the next step of the process consists on detecting objects in such region. For that purpose, we decided to follow a bottom-up approach. Instead of detecting specific objects, our system extracts basic features that could be used to describe every object: the shape and the colour. These low level features or characteristics could be presented to the human, using speech or any other communication channel, in order to improve de multimodal interaction process, and hence the human could select the characteristics of the object to interact with. Moreover, the human could teach the robot about the object it is detecting. For instance, if the robot detects a big orange round-shaped object the human is pointing at, then the human could tell the robot that this object is a basketball ball.

Shape-based features

Three basic shapes are used to define the objects of the environment: squares, circles and triangles. In order to detect these shapes, the input region of interest of the images is pre-processed following the next steps:

1) Convert the colour image to greyscale
2) Noise reduction
3) Border detection
4) Gradient operator
5) Candidate contour detection

All these steps are performed by using the open source computer vision library OpenCV [16], which includes implementations for all these methods. Once the contours have been extracted, the next step consists on detecting the shape of these contours. This was done by studying the points that are included in the contour and the convexity, using the appropriate OpenCV functions (cvApproxPoly and cvMinEnclosingCircle concretely). Each candidate contour will be labelled as square, circle or triangle or directly discarded if it don't pass any shape test.

Colour-based features

After detecting the shapes of the candidate objects, the system extracts colour information for each circle, square or triangle recognized in the scene. In order to be more robust to lighting variations, we used the YCbCr colour space (Fig. 5) instead of the standard RGB. Y stands for the luminance component and Cb and Cr are the chrominance components. Lighting variations involve higher variations for the luma component than for the chrominance ones, which makes easier to detect similar colours under different lighting conditions.

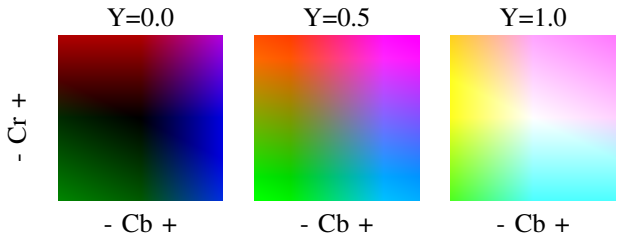

Fig. 5. YCbCr colour space.

The average colour computed for each shape is compared with 8 primary colours that will be used to define the object. These colours and their YcbCr and RGB values are summarized in Table I.

TABLE I
PRIMARY COLOURS USED TO DESCRIBE THE DETECTED OBJECTS.

Colour	YCbCr	RGB	sample
Red	$< 40, 109, 239 >$	$< 255, 0, 0 >$	
Black	$< 16, 128, 128 >$	$< 0, 0, 0 >$	
Green	$< 144, 34, 53 >$	$< 0, 255, 0 >$	
White	$< 235, 128, 128 >$	$< 255, 255, 255 >$	
Blue	$< 81, 239, 90 >$	$< 0, 0, 255 >$	
Yellow	$< 169, 16, 165 >$	$< 255, 255, 0 >$	
Orange	$< 105, 62, 202 >$	$< 255, 128, 0 >$	
Purple	$< 94, 230, 146 >$	$< 128, 0, 255 >$	

The difference between two colours is computed by using Eq. (3), where $\Delta(Y)$ denotes the absolute difference between the luma component of the two colours to compare. This equation is based on the original ΔE measurement function, proposed by the International Commission on Illumination[3]. The ΔE function was defined for being used with the Lab colour space, which also uses L for lightness and a and b for chrominance. The original ΔE value determines the quality of the comparison between two colours:

- ΔE values between 0 and 1: Excellent
- ΔE values between 1 and 2: Good
- ΔE values between 2 and 4: Normal
- ΔE values between 4 and 5: Low
- ΔE values higher than 5: Bad

In this work, we don't want to estimate the quality of the colour comparisons but use reliable colour distance measurements. Therefore, we use the $\Delta E'$ (see Eq. (3)) function to estimate the distance to the 8 primary colours. The colour under study will be labelled with the primary colour that obtained the lower difference.

$$\Delta E' = \sqrt{\Delta(Y)^2 + \Delta(Cb)^2 + \Delta(Cr)^2} \qquad (3)$$

[3]http://www.cie.co.at/

Thanks to the use of the shape-based and colour-based features, the robot would be capable of describing the scene in a way comprehensive to humans. These low level features are inherently ambiguous and several elements of the environment can share them. However, the ambiguity is reduced because only the part of the scene that the human is pointing at is processed.

V. EXPERIMENTS

The main objective of the experiments is to evaluate the performance of the system. Concretely, we want to demonstrate that a) it is possible to detect when a user is requiring the interaction with an object and b) the robot is capable of providing information about the candidate objects for the interaction.

The first experiment was focused on the generation of the classifier. It showed all the steps performed to obtain the final classifier, which uses the depth information as input. The second experiment evaluated the extraction of depth-based and colour-based features from visual images.

We used the following parameters for all the experiments:

- Computer: AMD Phenom 2x6 1055T, 2.8GHz and 4GB RAM memory.
- Depth and visual Camera: Microsoft Kinect device.
- Classification algorithm: C4.5 within Weka toolkit.
- Environment: 265x400 squared indoor room
- Lighting conditions: 50hz light bulb
- Distance between human and camera: 2-4 metres
- Height of the camera: 90 cm
- Intrinsic parameters of the Kinect device: standard

A. Experiment 1.-Classifier

The objective of the classifier was to estimate the pose of the user from the data provided by the Kinect depth sensor. This data consisted of the position of several parts of the body that were obtained with the OpenNI framework: the head, the neck, the bust and pair of values (right and left) for the hands, the elbows and the shoulders. For all these 9 points, OpenNI detect their $< x, y, z >$ absolute position in the world. In addition to the position of these points, OpenNI also provided the orientation of the head. As mentioned before, three different classes (or poses) were used: "NoAttention", "AttentionLeft" and "AttentionRight".

We generated a balanced dataset by recording a human gesturing in the selected environments. This dataset consisted of 447 samples, where each consist of 28 different features. The original dataset was processed to evaluate the effect of modifying the reference system. Our proposal was to store the distance to a reference point (the neck) instead of the absolute position for the hands or the elbows. This modification was expected to improve the generalization of the classifier, trying to avoid bias and overfitting problems.

We generated two decision trees (C4.5) for the original and the modified datasets. The accuracy classfication rate was 82.4% for the original dataset and 85.6% for the second one. Despite of this improvement was not significant, we opted for using the modified dataset because of the more capability for generalization and, in addition, because the reduced number of features it uses (the position of the neck was discarded).

The second step of this experiment focused on a feature selection step. We tested the "InfoGaintAtributeEval" evaluator along with the "Ranker" method in Weka, and this selection obtained a dataset with 13 features and a accuracy classification rate of 89.5%. We also tested the "Wrapper" evaluator, which obtained the best results: 89.5% of accuracy classification rate with only 11 features. All the results concerning the classifier are shown in Table II. According to these results similar performance was obtained using the the two studied feature selection algorithms, but the "Wrapper" obtained the lowest number of useful features.

TABLE II
SUCCESS RATES FOR THE HUMAN POSE CLASSIFIER.

Dataset	Feature Selection	Success Rate	#Features
Original	None	82.4%	28
Modified	None	85.6%	25
Modified	InfoGaintAtributeEval	89.5%	13
Modified	Wrapper	89.5%	11

B. Experiment 2.- Shape-based and colour-based features extraction

1) Shape-based features: The first objective of the second experiment was to evaluate the detection of the objects in a region of interest. We applied all the processing already shown in Section IV-B1, trying to detect all the circles, squares and triangles. The scenario for the test (Fig. 6) was a white wall with several objects: 6 triangles, 5 squares and 4 circles.

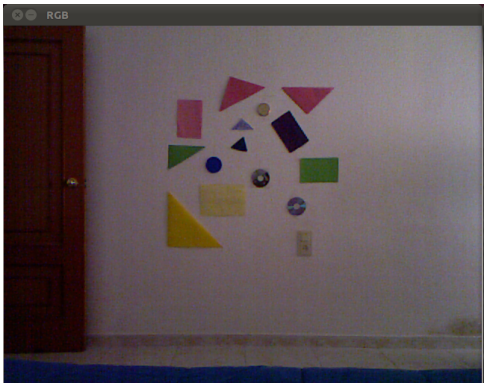

Fig. 6. Objects to extract their features.

We processed 30 visual images acquired in this environment. The classification results for the proposed algorithm and all the shapes is summarized in Table III.

TABLE III
SUCCESS RATE FOR THE SHAPE-BASED FEATURES DETECTION.

Object	Detected	Not Detected	Success Rate
Triangle	125	55	66.44%
Circle	117	3	97.5%
Square	113	37	75.33%

2) Colour-based features: In this experiment we testes the performance of the object colour estimation algorithm. We used the same scenario as in the previous experiment and we compared the colour detected for each object with the real one. The real colours of the objects in the scenario were manually labelled.

The colour detector consisted on classifying each object with the most similar primary colour. It was done by using the $\Delta E'$ formula (Eq. (3)), the YCbCr colour space and 8 primary colours: red, black, green, white, blue, yellow, orange and purple. Table IV shows the classificationn accuracy of the colour-shape features detection separately for the colours that appear in the dataset.

TABLE IV
SUCCESS RATE FOR THE COLOUR-BASED FEATURES DETECTION.

Colour	Detected	Not Detected	Success Rate
Green	57	3	95.0.%
Yellow	41	11	78.86%
Pink/Purple	85	5	94.44%
Black	30	30	100.00%
White	9	19	32.14%
Blue	63	27	70.00%

Despite the results shown in Tables III and IV could be improved by using more sophisticated techniques (classification and feature extraction algorithms) they can be considered useful for the purpose of this work.

Fig. 7 shows an image where all the process is applied. Firstly, the pose classifier detected that the human wanted to interact with the objects located at the left of the image (using the depth information). After that, the system selected the region of interest and detected 5 different objects using their shape. Finally, for each one of these objects the system estimated its colour.

Fig. 7. Example of the system.

It can be observed how the shape-based feature extraction failed on the detection of a small triangle. Moreover, not all the colours were correctly classified: dark blue was labelled as black and the white circle as green.

In the light of the experimental results obtained, we assume that for a final real scenario, with more sophisticated objects to detect and not idel scenarios, more investigation have to be performed in order to improve the object detection and classification algorithms. Nevertheless, the obtained results are acceptable for the final purpose of this work, which is to stablish a starting point to stablish a communication between humans and robots in order to achieve an agreement of what objects one or both of them want to interact with.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a visual attention system based on the use of depth and visual information. This system has been proposed for providing robots with the capability of detecting when a user wants to interact with an object of the environment.

For this purpose, we have developed a human pose estimation system and a low level feature extractor. The pose estimation system is used to select a region of interest from the visual image, where all the features are extracted. These low level features (shape-based and colour-based) can then be used to detect and recognize different type of objects in the region of interest, and therefore can be helpful to start a conversation with a human, acting in anticipation and asking about the objects to interact with. Nevetheless, more investigations are needed in order to improve the recognition/classification results and to deal with more sophisticated objects and environments.

For future work, we have in mind to use this development as a starting point for a complete multimodal human robot interaction system. We also have plans to develop a cognitive representation model for describing the objects of the environment.

ACKNOWLEDGEMENTS

This work has been partially supported by the European Social Fund, FEDER, Spanish Ministerio de Ciencia e Innovacin (MICINN) and Junta de Comunidades de Castilla-La Mancha regional government under TIN2010-20900-C04-03, PBI08-0210-7127 and PPII11-0309-6935 projects.

REFERENCES

[1] T. Fong, I. R. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[2] G. Littlewort, M. Bartlett, I. Fasel, J. Chenu, T. Kanda, H. Ishiguro, and J. Movellan, "Towards social robots: Automatic evaluation of human-robot interaction by face detection and expression classification," in *Advances in neural information processing systems*, S. Thrun, L. Saul, and B. S. (Eds.), Eds. Cambridge, MA: MIT Press, 2004, vol. 16, pp. 1563–1570.

[3] H. Okuno, K. Nakadai, and H. Kitano, "Social interaction of humanoid robot based on audio-visual tracking," in *Developments in Applied Artificial Intelligence*, ser. Lecture Notes in Computer Science. Springer-Verlag, 2002, vol. 2358, pp. 140–173.

[4] S. Li, B. Wrede, and G. Sagerer, "A computational model of multi-modal grounding for human robot interaction," in *SigDIAL '06: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 153–160.

[5] T. K. Harris, S. Banerjee, and A. I. Rudnicky, "Heterogeneous multi-robot dialogues for search tasks," in *AAAI Spring Symposium: Dialogical Robots*, 2005.

[6] D. Bohus, A. Raux, T. K. Harris, M. Eskenazi, and A. I. Rudnicky, "Olympus: an open-source framework for conversational spoken language interface research," in *NAACL-HLT '07: Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*. Morristown, NJ, USA: Association for Computational Linguistics, 2007, pp. 32–39.

[7] M. A. Goodrich and A. C. Schultz, "Human-robot interaction: A survey," *Foundations and Trends® in HumanComputer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[8] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot," in *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 2003, pp. 28–35.

[9] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke, "Multimodal conversation between a humanoid robot and multiple persons," in *Proc. of the Workshop on Modular Construction of Humanlike Intelligence at the Twentieth National Conferences on Artificial Intelligence (AAAI)*, 2005.

[10] R. Munoz-Salinas, E. Aguirre, M. García-Silvente, and A. González, "A fuzzy system for visual detection of interest in human-robot interaction," in *2nd International Conference on Machine Intelligence (ACIDCA-ICMI 2005)*, 2005, pp. 574–581.

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

[12] F. Orabona, G. Metta, and G. Sandini, "Object-based visual attention: a model for a behaving robot," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops - Volume 03*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 89–. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.502

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

[14] S. Vasudevan, S. Gächter, V. Nguyen, and R. Siegwart, "Cognitive maps for mobile robots – an object based approach," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 359–371, 2007.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[16] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.