Visual Perception System for an Interactive Robot

Rebeca Marfil, Juan P. Bandera, Antonio Bandera and Antonio Palomino
Departamento de Tecnología Electrónica
Málaga University
Campus de Teatinos 29071-Málaga (Spain)
{rebeca, jpbandera, ajbandera, ajpalomino}@uma.es

Ricardo Vázquez-Martín
CITIC
Parque Tecnológico de Andalucía
rvazquez@citic.es

***Abstract*— A key area in robotics research is concerned with developing social robots for assisting people in everyday tasks. A social robot is a robot which is capable, not only to navigate in unknown environments, but also to interact with people and to react to changes in its environment in an intuitive manner at the same time that they are performing other tasks. To achieve this intuitive interaction with people and with the environment, it is interesting for the robot to be able to perceive the real world in a similar way that people do, extracting from the sensed data a reduced set of relevant items. In this proposal, we focus on the study and develop of the visual perception system of the robot. The central part of this system will be an attention mechanism which must be able to discriminate, from all the visual information provided by the robots sensors, the most relevant elements needed to carry out the currently executed tasks. Specifically, two behaviours or tasks have been implemented. These behaviours employ the perception system to detect distinguished visual landmarks in an initially unknown environment, or to detect and capture the upper-body motion of people interested in interact with the robot. Both behaviours can be simultaneously conducted and they will allow the robot to perceive the surrounding environment.**

I. Introduction

People display a remarkably robust ability to extract the relevant information from the environment. Developing computational perception systems that provide these same sorts of abilities is a critical step in designing robots that are able to cooperate with people as capable partners, that are able to learn from natural human instruction, and that are intuitive and engaging for people to interact with, but that are also able to simultaneously navigate in initially unknown environments or to grasp a specific object.

Our proposal is focused on the study and development of the visual perception system for a social robot. A social robot is an autonomous robot which is not only able to navigate and solve common tasks, but it is also simultaneously able to communicate and interact with people and other social robots. This implies that the social robot should perceive a great variety of natural social cues from visual and auditory channels, and must reply to these stimulus at human rates. The proposed perception system will be used to extract from the input data the low-level information that the robot will need to accomplish the navigation and human-robot interaction tasks. Besides, to achieve an intuitive interaction with people, it is interesting that the robot will be able to perceive the real world in a similar way that people do. Thus, the socially interactive robot will interpret the same phenomena that people observe [1]. The central element of our proposal is an artificial attention mechanism which will be able to discriminate, from all the low-level information provided by the robots sensors, the most relevant data useful to fulfil the currently executed tasks and to detect people to interact with. The proposed attentional mechanism has two main contributions: i) While other methods associate a saliency value to each image pixel, our proposal computes a set of 'preattentive objects' or 'proto-objcets' with an associated saliency value; and ii) The inclusion of a semiattentive stage which will take into account the currently executed tasks in the information selection process. Thus, the importance of a set of sensed data will not only depend on low-level task-independent features but also on the tasks to reach. This semiattentive stage also implements a mechanism to avoid the revisiting of previously analyzed areas which resembles the so-called 'inhibition of return'.

II. Overview of the proposed attention mechanism

The proposed object-based model of visual attention integrates bottom-up (data-driven) and top-down (model-driven) processing. The bottom-up component determines and selects salient 'preattentive objects' by integrating different features into the same hierarchical structure. These 'preattentive objects' or 'proto-objects' [2] are image entities which do not necessarily correspond with a recognizable object, although they posses some of the characteristics of objects. Thus, it can be considered that they are the result of the initial segmentation of the image input into candidate objects (i.e. grouping together those input pixels which are likely to correspond to parts of the same object in the real world, separately from those which are likely to belong to other objects). This is the main contribution of the proposed approach, as it is able to group the image pixels into entities which can be considered as *segmented perceptual units* [2]. On the other hand, the top-down component could make use of object templates to filter out data and shift the attention to objects which are relevant to accomplish the current tasks to reach. Finally, in a dynamic scenario, the locations and shapes of the objects may change due to motion and minor illumination differences between consecutive acquired images. In order to deal with these scenes, a tracking approach for 'inhibition of return' is employed. To support the tracking process, the computed hierarchical representations will include recently proposed

templates associated to the appearance and motion of tracked items [3]. All processes will be conducted using the same hierarchical structure: the Bounded Irregular Pyramid (BIP) [3,4]. Its application to this framework is the second main novelty of the proposed model.

The proposed attention mechanism consists of three main modules. The first one implements a concept of saliency based on 'preattentive objects' [2]. From these objects, different saliency maps are generated. These maps are the input of the semi-attentive stage [5,6]. At this stage, significant items according to the tasks to accomplish are identified and tracked. This tracking process allows to implement the 'inhibition of return', avoiding to detect these items as new significant 'preattentive objects'. The attentive stage fixes the field of attention to the most salient object depending on the current behaviour. It must be noted that our 'inhibition of return' is not exactly equal to the one described for human beings. In our case, it allows only focusing on non-visited image regions.

In order to provide to a social robot with the necessary abilities to autonomous navigate and interact with people in a dynamic scenario, two main behaviours have been included in the proposed attentional mechanism: a human gesture recognition behaviour and a visual natural landmark detection one. These behaviours are the responsible to recognize a person who is interested in establishing an interaction, and to provide visual natural landmarks for mobile robot navigation, respectively. To achieve this, the attentive stage will take into account both static items environment visual landmarks- and dynamic ones human body parts- which could be used by other high-level modules of the robot control architecture.

III. Visual landmark detection task

As it was aforementioned, the preattentive stage provides a set of saliency maps. Among the set of regions which constitute these maps, the visual landmark detection behaviour selects those which satisfy certain conditions. The key idea is to use as landmarks rectangular shaped regions or quasi-rectangular shaped regions without holes. In this way, we try to avoid the selection of segmentation artifacts, assuming that a rectangular region has less probability to be a segmentation error than a sparse region with a complex shape. Selected regions cannot be located at the image border in order to avoid errors due to partial occlusions. On the other hand, in order to assure that the regions are almost planar, regions which present abrupt depth changes inside them are also discarded. Besides, it is assumed that large regions could be more likely associated to non-planar surfaces. Finally, The selected regions must also exhibit a relatively high contrast with respect to its surroundings.

Extensive tests have shown that this detector reliably finds the same visual features under different viewing and illumination conditions [7].

IV. Human gesture recognition task

In order to identify the movements of a human which is interacting with the robot, this behaviour selects from the whole set of regions provide by the semiattentive stage of the attention mechanism those which correspond with the face and hands of the human. These regions are used to track and recognize the movements of the person using novel human motion capture [8] and gesture recognition approaches [9].

V. Conclusions

Experimental results reveal that the object-based attention mechanism has been successfully integrated with an attentive stage which control the field of attention following two different behaviours: a human gesture recognition behaviour and a visual landmark detection one. If previously published works [7-8-9] mainly dealt with the implemented behaviours, this work is focused on the attention mechanism and its ability to merge bottom-up and top-down components of visual attention.

Acknowledgment

This work has been partially granted by the Spanish Ministerio de Ciencia e Innovación (MICINN) and FEDER funds, Project n. TIN2008-06196 and by the Junta de Andalucía, Project n. P07-TIC-03106.

References

[1] K. Dautenhahn, & C. Nehaniv, *Imitation in animals and artifacts*, Cambridge, MA: MIT Press, 2002.

[2] F. Orabona, G. Metta,& G. Sandini,A proto-object based visual attention model, In L. Paletta and E. Rome (eds.) *WAPCV 2007*, LNAI **4840**, pp. 198–215, Heidelberg: Springer, 2007.

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

[4] R. Marfil, L. Molina-Tanco, A. Bandera, & F. Sandoval, The construction of bounded irregular pyramids using a union-find decimation process, In F. Escolano and M. Vento (eds.) *GbRPR 2007*, LNCS **4538**, pp. 307–318, Heidelberg: Springer, 2007.

[5] G. Backer, & B. Mertsching, Two selection stages provide efficient object-based attentional control for dynamic vision, *WAPCV 2003*, pp. 9–16, Heidelberg: Springer, 2003.

[6] R. Marfil, A. Bandera, J.A. Rodríguez & F. Sandoval, A novel hierarchical framework for object-based visual attention, In L. Paletta and J.K. Tsotsos (eds.) *WAPCV 2008* LNAI **5395**, pp. 27–40, Heidelberg: Springer, 2009.

[7] R. Vázquez-Martín, R. Marfil, P. Núñez, A. Bandera, & F. Sandoval, A novel approach for salient image regions detection and description *Pattern Recognition Letters* **30** (16): 1464–1476, 2009.

[8] J.P Bandera, R. Marfil, J.A. Rodríguez, L. Molina-Tanco & F.Sandoval, A novel hybrid approach to upper-body Human Motion Capture, In 14th IEEE Mediterranean electrotechnical conference (Melecom 2008), 355-360, 2008.

[9] J.P. Bandera, R. Marfil, A. Bandera, J.A. Rodríguez, L. Molina-Tanco, & F. Sandoval, Fast gesture recognition based on a two-level representation *Pattern Recognition Letters* **30** (13): 1181–1189, 2009.