Bootstrapping Visual Memories for a Humanoid Robot

Giovanni Saponaro Alexandre Bernardino
Institute for Systems and Robotics, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisbon, Portugal
{gsaponaro,alex}@isr.ist.utl.pt

Abstract

This paper addresses the problem of creating visual memories of salient objects arising in a certain unknown environment and the ability of recalling them at a later time. We adopt a developmental approach, i.e., we start with low complexity and coarse information that is incrementally refined to create a hierarchy of visual representations for the observed scenes. Colour histogram-based measurements provide a simple and quick methodology to bootstrap the creation of visual memories. Experiments performed on the humanoid platform iCub illustrate how this technique can successfully create concepts by grouping related images together in clusters, subject to a similarity threshold.

Figure 1. Experiment with threshold = 0.8, with two of the obtained clusters displayed.

1. Introduction

Personal Robotic Assistants will be deployed in public and private settings in the near future, having to deal with dynamic and uncertain information. In particular, they must perceive –visually or by other means– the relevant objects and places of their environment. While some amount of preprogramming can be added to provide these robots with the ability to detect certain common objects (e.g., faces, cars, doors, tables), it is not possible to predict all the situations that the robot will have to deal with in complex conditions. Thus, developmental robotics approaches are studying the problem of how robots can incrementally acquire knowledge, relying solely on observations and self-experiences obtained *in situ* [2].

Our work is developed on the humanoid robot platform iCub[1], where the specification of an Auto-Associative Memory (Episodic Memory) is under development [2]. Such a memory operates as follows: (i) when an image is presented to the memory, it attempts to recall that image with vision algorithms; (ii) if a previously-stored image matches the presented image sufficiently well, then the stored image is recalled and displayed; (iii) if no previously-stored image matches sufficiently well, then the presented image is stored in the database.

The whole process is controlled by a parameter that establishes the required matching degree to associate two images.

In this paper we study how this simple methodology can provide desirable properties in the creation and remembrance of relevant visual memories. We consider that the robot does not know about anything at first. It is then attracted by highly-salient parts of the scene using bottom-up attentional mechanisms with colour filters [3]. Due to the initial lack of knowledge, the system must create a small, non-specific set of visual memory classes, in order to limit system complexity. With time, once "first degree" memories are stable, the system must start distinguishing from images of the same class and incrementally create less abstract representation of objects. We study how controlling the

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

[2] http://eris.liralab.it/wiki/Auto-associative_Memory_Specification, written by D. Vernon [5].

matching degree parameter of the above-mentioned Auto-Associative Memory influences the size of the created memories and the ability of developmentally growing hierarchical representations.

2. Related Work

A powerful algorithm used in visual object recognition is Scale-Invariant Feature Transform (SIFT). Figueira [1] applied it to cognitive robotics, implementing the spatial model surrounding a humanoid iCub robot with the aim of identifying salient objects which the robot encounters and memorizes during its visual exploration, treating said objects as clusters of SIFT features.

However, using SIFTs to address the recognition task of objects has two shortcomings. First, for objects to be effectively characterized and recognized with this method, they need to have highly-textured external surfaces (for instance, a uniformly yellow cuboid would not be correctly matched, but a package of breakfast cereal would). Secondly, another assumption is that the facets of the objects to treat be planar and non-rigid; so, the head of a person yawing in front of the camera from various perspectives would not be correctly matched, despite having a textured surface.

Colour histogram-based algorithms, on the other hand, are flexible but not very selective, resulting in abstract categorization: the two images in the bottom row of Fig. 1 are considered as members of the same class, even though they depict different people.

3. Proposed Approach

To compare a pair of images in the visual memory, we use a Histogram Intersection technique [4] that returns the value

$$H(I, M) = \frac{\sum_{j=1}^{n} \min(I_j, M_j)}{\sum_{j=1}^{n} M_j} \qquad (1)$$

where I and M are the two histograms ("input" and candidate "model"), each with n bins. A histogram intersection $\sum_{j=1}^{n} \min(I_j, M_j)$ is defined as the number of pixels of the same colour in the image. Because of the division by the number of pixel in the M histogram, the value of H is normalized between 0 and 1.

Within this framework we will study two main issues:

1. which saliency criteria to use for the selection of candidate images to store in the visual memory: since we are using colour-based distance metrics, highly saturated colours imply high salience;

2. how a single threshold on the categorization distance can be modulated to create hierarchical class representations: in order to form a new class, an image must be distinct from all the other images that are already present in the memory.

As far as computational complexity of histogram intersection is concerned, it is linear in the number of elements in the histograms. In our current implementation we only consider the Hue and Saturation components in HSV/HSI colour image representations, resulting in 16×16 histogram bins that are computed nearly in real time on a modern computer that receives images from the iCub cameras (Point-Grey Dragonfly 2, 640×480 pixels, 30 frames per second).

4. Results and Conclusions

An evaluation of the performance allows us to observe different behaviours shown by the classifier: a low threshold parameter implies generalized object preference (few new images are saved, i.e., at the end of exploration experiments there is a low number of clusters). On the other hand, if the threshold grows towards 1.0, a specific object preference is shown by the robot memory (high number of saved images, i.e., high number of clusters with unary cardinality).

Figure 2. Exploring an empty laboratory with threshold = 0.7 resulted in 5 image clusters being saved before stabilizing the system. Such a low number of classes was expected, given the rather uniform colours in the environment.

Table 1. Number of image clusters saved while exploring an empty laboratory environment, with varying values of the threshold.

threshold	≤ 0.4	0.5	0.6	0.7	0.8	0.9	0.95
# images	1	2	3	5	8	9	40

Figure 3. Cluster of plants obtained with a threshold value of 0.5.

Figure 4. Cluster of faces, all grouped together, obtained with a threshold value of 0.5.

Initially, we let the iCub explore its everyday environment, an empty laboratory, with different threshold values. Results in terms of number of saved image clusters that were obtained are listed in Table 1, while Fig. 2 shows all the classes acquired in one of these tests.

Other experiments consisted of several objects, people and scenarios presented in front of the iCub head. Each time a sufficiently salient and distinct image was detected, a new class was added to the memory. In order to obtain sensible clustering, we needed to use a rather low threshold, as shown is Fig. 3 and Fig. 4. Failure to do so –that is, using a higher threshold– would create too many distinct image classes that in reality represent the same concept, which we want to capture with our cognitive memory.

To conclude, we propose a technique to implement visual cognitive memories. The presented results show how the system can successfully create concepts by grouping related images together. Most importantly, it can do so in real time and starting from zero built-in knowledge. However, initializing and controlling the threshold parameter is critical, and this should be subject to future investigation.

Furthermore, issues like keeping memory size within limits with very large image databases, implementing a multi-layer hierarchical representation, controlling the branching factor, and testing different visual saliency approaches will be subject to future work.

5. Acknowledgements

This work was supported by Fundação para a Ciência e a Tecnologia (ISR/IST pluriannual funding).

References

[1] D. Figueira, M. Lopes, R. Ventura, and J. Ruesch. From Pixels to Objects: Enabling a Spatial Model for Humanoid Social Robots. In *IEEE International Conference on Robotics and Automation (ICRA 2009)*, Kobe, Japan, May 2009.

[2] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental Robotics: A Survey. *Connection Science*, 15(4):151–190, Dec. 2003.

[3] J. Ruesch, M. Lopes, A. Bernardino, J. Hörnstein, J. Santos-Victor, and R. Pfeifer. Multimodal Saliency-Based Bottom-Up Attention: A Framework for the Humanoid Robot iCub. In *IEEE International Conference on Robotics and Automation (ICRA 2008)*, Pasadena, CA, USA, May 2008.

[4] M. H. Swain and D. H. Ballard. Color Indexing. *Internation Journal of Computer Vision*, 7(1):11–32, Nov. 1991.

[5] D. Vernon. Deliverable D2.1 – A Roadmap for the Development of Cognitive Capabilities in Humanoid Robots, Oct. 2008. `http://www.robotcub.org/index.php/robotcub/more_information/deliverables`.