

How Are Representations Affected by Scene Statistics in an Adaptive Active Vision System?

Dimitri Ognibene*

Giovanni Pezzulo**

Gianluca Baldassarre*

*ISTC-CNR Via S.Martino della Battaglia, 44 - 00185 Rome, Italy

**ILC-CNR Via Giuseppe Moruzzi, 1 - 56124 Pisa, Italy

{dimitri.ognibene,giovanni.pezzulo,gianluca.baldassarre}@istc.cnr.it

1. Introduction

One of the main claims of active vision (Ballard, 1991) is that finding data on demand, based on the requirements of the task, is more efficient than reconstructing the whole scene by performing a complete visual scan of it. This aids generalisation and a dramatic reduction of the needed visual computations. Using this strategy, however, generates the need to learn complex gaze control strategies dependent on the pursued goals and the properties of scenes and objects. For example, to be able to find an object in the environment an agent needs to learn to use several sources of information such as spatial relations of objects and bottom-up saliency of scene regions. In addition, if the system is genuinely autonomous it also needs to develop a representation of the objects themselves, for example of potential targets, cues and distractors, on the basis of generic reward signals to be maximized and the visual control policy used. Most of the models proposed in developmental robotics do not use *adaptive* visual control and so are ill suited to investigate these issues.

In a previous work (Ognibene et al., 2008) we presented a reinforcement-learning neuro-robotic architecture, based on neural population codes, which was able to *develop attention control* policies by interacting with the environment *based on a rewarded reaching task* it had to accomplish. In this paper the same architecture is used to investigate the types of *internal representations* that this same architecture develops when exposed to two classes of environments where objects are organised on the basis of *contrasting spatial relations* (Figure 1).

A recent view on neural population codes proposes that neural maps might be used to develop overall probability distributions of stimuli (Pouget et al., 2002). On the contrary, this study shows that active vision systems tend to develop actions which actively disambiguate the stimuli and acquire new evidence only when needed: as a consequence, the acquired representations do not reflect overall probability distributions related to stimuli but rather the contextual relationships between them.



Figure 1: Examples of environments used to test the model, drawn from two classes of environments **L** and **R**. In each trial, the specific environment was randomly drawn from **L** or **R** with a probability of 75% and 25%, respectively. Both classes of environments were based on 2 to 5 green cues forming a vertical line, one blue distractor, and one red target. The cues, distractor and target were located on the vertexes of a 5×6 matrix. In **L** environments, the target and distractor were located at a random position respectively at the left and right of the green line, whereas in **R** environments were located at a random position respectively at the right and left of it.

2. The model

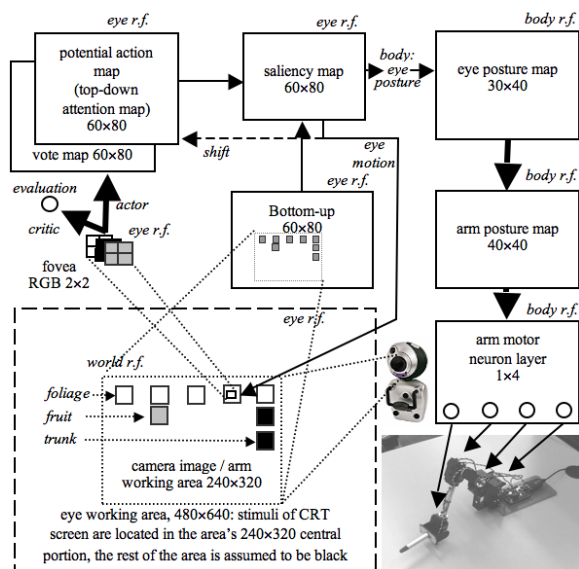


Figure 2: The architecture of the model.

The architecture and setup of the model (Figure 2.), used here in a simulated version, is now

briefly described but the reader should refer to Ognibene et al. (2008) for details. The robotic setup used to test the model is composed of a camera looking down to a robotic arm. The arm acts on a working plane consisting of a screen which shows the visual stimuli of the task.

The architecture of the model is formed by three main components:

(a) *Bottom-up attention component.* The input image is used to activate a *periphery map* which identifies high-contrast regions on the basis of suitable filters .

(b) *Top-down attention component.* The central part of input image (*fovea*) is used as input of an *actor-critic* model which learns to predict, by suitably activating the output map of the actor (*vote map*), the spatial position of the rewarded targets with respect to the foveated objects. A *potential action map (PAM)*, based on leaky neurons, accumulates evidence, furnished by the actor, on possible locations of the target while the fovea explores the scene objects. An overall *saliency map* integrates information from the periphery map and the PAM to select the next eye movement on the basis of a dynamic neural competition. All maps of the attention components use an eye-centered reference frame.

(c) *Arm-control component.* Each fixation point, encoded in a *eye posture map*, suggests a potential target to a *arm posture map*: when the eye fixates a location for enough time (3 time steps on average), the arm posture map triggers a related arm action on the basis of a second dynamic neural competition. If the reached object is the target, the system gets a reward of one, otherwise it gets a small punishment (mimicking energy consumption).

3. Results and Conclusions

The tests of the model show that it learns an exploration policy which initially assumes to be tackling an **L** environment, so first searches the green line and then, on this basis, the target on its left (two eye steps). In the presence of an **R** environment, this assumption fails and the agent searches the target directly on the right of the green line rather than exploring anew. This strategy allows the system to find the target with only one additional step.

Table 1 shows the activation of the vote map of two agents respectively trained with **L** environments or with both **L** and **R** environments (with a frequency of 75% and 25%, respectively), when the agents foveate either the cue or the distractor (a third agent trained only with **R** environments developed vote maps mirroring those of the **L**-trained agent: data not reported).

These results show that the representations underlying the gaze-control policies are not based on a combination of all possible policies needed to tackle

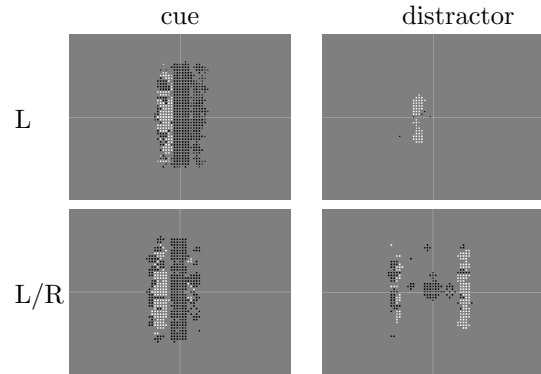


Table 1: Activation of the vote map when the model foveates the cue or distractor. L: agent trained only on **L** environments for 60.000 trials. L/R: agent trained on both **L** and **R** environments for 60.000 trials.

the two classes of environments. In fact in the latter case one would expect internal representations to be a combination of the vote maps needed to tackle the **L** or **R** environments in isolation (e.g., a sum or a max of the two). Instead, the internal representations encode the specific exploration routines best suited to solve the task at hand. This is especially evident if one considers the vote maps related to the distractor: when the system is trained with **L** environments, the map does not encode the position of target but only the action of foveating the green line, whereas when trained with both **L** and **R** environments the system encodes the action of going to the right of the green line as in this case the distractor becomes a good predictor of the target located there.

These strategies exemplify a general principle used by adaptive active vision system to tackle complex environments. When agents must learn to autonomously and adaptively solve tasks, the representations they develop reflect the actions that permit to interact with the environment in order to acquire new information and solve tasks given the information acquired that far, more than the overall statistics of scenes.

Acknowledgements Research funded by the EU project IM-CLeVeR (FP7-ICT-IP-231722).

References

- Ballard, D. (1991). Animate vision. *AI*, 48:57–86.
- Ognibene, D., Balkenius, C., and Baldassarre, G. (2008). Integrating epistemic action (active vision) and pragmatic action (reaching): A neural architecture for camera-arm robots. In *SAB’08*, Osaka, Japan. Springer.
- Pouget, A., Ducom, J. C., Torri, J., and Bavelier, D. (2002). Multisensory spatial representations in eye-centered coordinates for reaching. *Cognition*, 83(1):B1–11.