Berthouze, L., Kaplan, F., Kozima, H., Yano, H., Konczak, J., Metta, G., Nadel, J., Sandini, G., Stojanov, G. and Balkenius, C. (Eds.)
Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems
Lund University Cognitive Studies, 123. ISBN 91-974741-4-2

Towards Teaching a Robot to Count Objects

Julien Vitay
Loria / INRIA
Campus Scientifique, B.P. 239
54506 Vandoeuvre-lès-Nancy Cedex, FRANCE
Julien.Vitay@loria.fr

Abstract

We present here an example of incremental learning between two computational models dealing with different modalities: a model allowing to switch spatial visual attention and a model allowing to learn the ordinal sequence of phonetical numbers. Their merging via a common reward signal allows anyway to produce a cardinal counting behaviour that can be implemented on a robot.

1. Context

The constructivist theory of learning (Piaget, 1972, Vygotsky, 1986) states that cognitive development relies on relatively discrete stages, where the infant learns new schemes on the basis of formerly acquired schemes in the previous stage. Two transitions between stages are of particular interest for the neurobotics community: the acquiring of sensorimotor schemes from motor reflexes; the acquiring of basic language abilities like semantics from sensorimotor schemes. In particular for the second transition, the quite recent discovery of the so-called "mirror-neurons" in the premotor area F5 of the monkey (Rizzolatti et al., 1996) (which respond equally for the execution and the observation of an action) has lead (Rizzolatti and Arbib, 1998) to explain the acquiring of language via the common abstract representation of sensorimotor schemes between the learner and his social environment.

As this indicates that the semantics of an action (either performed or recognized) is linked to its motor preparation, the same seems to be true with the semantics of an object. The sensorimotor contingency theory by (O'Regan and Noë, 2001) states that seeing is not building an internal representation of the whole visual information but rather exploring via visuomotor schemes (for example saccades) the behaviourally relevant location and ignoring the others. A striking evidence is given by the "change blindness" experiments which showed how the disappearance of a massive part of an image can be totally unreported by a subject if this part were not relevant for the understanding of the scene.

This idea of using previously acquired sensorimotor schemes to learn the semantics of an action or an object is in our view the major issue in autonomous robotics: the work done by Aude Billard (Schaal et al., 2003), Luc Steels (Steels, 2003) and Jun Tani (Sugita and Tani, 2002) for example enlights the advantages of that approach compared to classical artificial intelligence (based only on explicit representations).

In this paper, we present an example of incremental learning of a cognitive ability (counting objects in a scene) using a previously acquired sensorimotor skill (switch of attention on salient targets). We will first briefly describe the proposed task and then present the two different models and their merging.

2. Numbering Objects

Interaction of a robot with its environment needs non-linear and complex computations to achieve a successful behaviour. In particular in natural scenes, targets are not always unique: the task "bring me three apples" does not specify which apples are to be brought. In such a task, a robot would have first to determine if there is enough apples in the scene: determining the size of a set is called cardinality. When performing the task itself, the robot has to know that the first apple is followed by the second one and then by the third: determining the position of an item in a sequence is called ordinality.

The relationship between these two aspects of numbering in developmental psychology is not yet clear (see (Brannon and Van de Walle, 2001) for a debate). Young infants (< 2 years) seem to have a cardinal ability limited to 3 or 4 items called "subitizing", but this ability only improves with the acquiring of verbal counting (with counting rhymes for example) at the age of 3.5 or 4 years. It is only when they master the verbal sequence "one two three four five.." that they are able to tell that four objects are less than five. In other words, they have to make the correspondance between the "four" word related to the verbal sequence and the four objects in front of them, what is a cross-modal task (between phonological inputs and visual properties).

In this paper we will not consider the subitizing

process and make the assumption that counting objects has to grow on two distinct but parallel abilities: the ability to sequentially focus the interesting objects in the visual scene and the ability to learn the counting sequence "one two three four...". In the two next sections, we will briefly present biologically-inspired computational models for these two abilities and then show how then can be coupled to produce cardinal comparisons.

3. A Spatial Attention-Switching Mechanism

There is no need in a robotic task to analyze every pixel in the image grabbed by the camera of the robot. Important targets or clues are salient locations on the image, with respect to different features (colour, orientation, luminosity, movement, etc.) and at different scales. Analysing a visual scene is then sequentially focusing these salient locations until the correct target is found. Furthermore, the features involved in the computation of salience can be biased by task requirements.

In (Vitay et al., 2005), we presented a computational model of dynamical attention-switching based on the Continuum Neural Field Theory (Amari, 1977, Taylor, 1999), a framework of dynamical lateral interactions within a neural map. Each neuron is described by a dynamical equation which asynchronously takes into account the activity of the neighbouring neurons via a "mexican-hat" lateral-weight function. We pinpointed some interesting properties of that framework in (Rougier and Vitay, 2005) like denoising and spatio-temporal stability. Combining different neural maps with the same dynamical equation while playing with afferent and lateral weights, we were then able to make emerge a sequential behaviour from this completely distributed substrate.

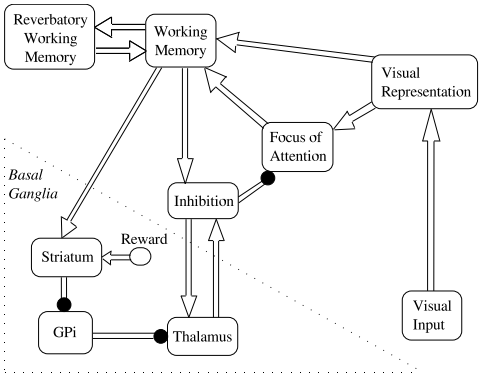

Figure 1: Attention Switching Architecture: empty arrows represent excitatory connections, round arrows represent inhibitory ones. For details, see (Vitay et al., 2005).

Figure 1 shows the architecture of the proposed system, but its description would need too much neurophysiological jargon for the audience of this workshop. We just need to say that it is composed of four sub-structures:

- a visual representation system fed by the saliency image that can filters out noise and allows only one salient location to be represented in the FOCUS OF ATTENTION map.

- a working memory system that enables to dynamically remember previously focused objects.

- an inhibition mechanism which can move the focus of attention to a new salient location (with the information given by the working memory).

- a basal-ganglia-like channel which can control the time of the switch of attention. The key signal is a phasic burst of dopaminergic activity in the REWARD unit.

As a consequence of this distributed and dynamical architecture, the serial behaviour that emerges is the sequential focusing of the different salient points on the image, without ever focusing twice the same object. Moreover, the time of the switch is controlled by the dopaminergic burst in the REWARD unit. We will use that property for the counting task presented later. This model has been successfully implemented on a *PeopleBot*® robot, whose task was to successively focus with its mobile camera a given number of green lemons. A nice feature of this model is that it can work either in the covert mode of attention (without eye movement) or in the overt mode (with eye movements, because of the dynamic updating of the working memory with visual information).

4. A Sequence Learning Mechanism

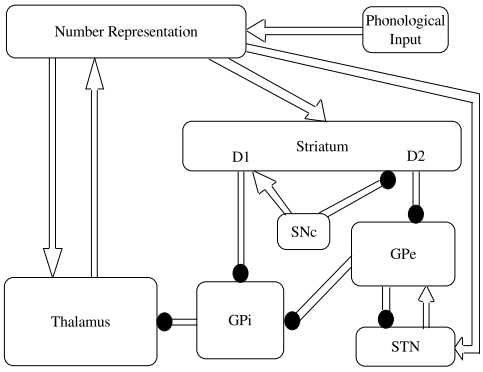

Figure 2: Ordinality Learning Architecture: empty arrows represent excitatory connections, round arrows represent inhibitory ones.

This system relies on the basal ganglia (BG) architecture, as summarized by (Hikosaka et al., 2000) and is directly inspired by the model made by (Berns and Sejnowski, 1998). BG are known to be

composed of several segregated channels, each of which being involved in a particular functional loop with the frontal cortex and its corresponding part of the thalamus to achieve selection of action. Basically, a thalamo-cortical network (bidirectional excitatory connections) is tonically inhibited by the BG output (here the internal segment of the Globus Pallidus GPi). The core mechanism of BG is disinhibition, which means that inhibition of the GPi (by the Striatum or by the external part of the Globus Pallidus GPe) disinhibits the corresponding part of the thalamus, thus allowing reciprocal excitation between the thalamus and the cortex. This mechanism has already been used in the attention-switching system.

Without giving too much detail, there are two opposite pathways in the BG architecture: the direct pathway that favorizes disinhibition and an indirect one which prenvents disinhibition. The balance between these two pathways is ensured by the dopaminergic signal produced by the Substancia Nigra pars compacta (SNc).

This complex architecture with bidirectional connections, internal loops and dopaminergic modulation allows to learn sequences and do selection of action (Berns and Sejnowski, 1998, Gurney et al., 2004). We will describe here how our model is able to learn simple sequences like ordinal numbers (cf. Figure 2).

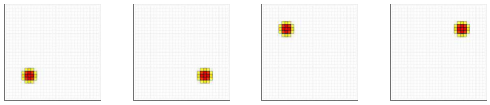

Figure 3: The distributed representation of the phonetical words ’zero’ ’one’ ’two’ ’three’ in the NUMBER MAP.

- We first organized a neural map called NUMBER representation map with phonological inputs representing numbers (zero, one, two, three). The hearing of one of these numbers therefore implies the appearance of a bubble of activity in the NUMBER map at a given location on the map (Figure 3). Please note that this coding is not compact and would not scale to large numerosities.

- We then present successively the four numbers to the BG model with a phasic burst of dopamine in SNC at the time of the switch. As dopamine stands for a kind of “reward” signal (Schultz et al., 1992), we can justify this by saying that hearing a voice (one’s mother’s voice for example) is intrinsically rewarding.

- The inner dynamics of the system (not described here) ensure that the association between the cortical representation of a number and its follower is learned by the connections between STN and GPE.

After learning, when some STN neurons are active for the current number (via their cortical inputs), they tend to excite GPE at the location of the next number, which in turn inhibits GPI. This artificially creates disinhibition in the thalamus that can be used to predict the next number.

- A high tonic level of dopamine favorizes the direct pathway so that the representation of the current number in the NUMBER map is stable even without any phonological input (it is mainly the same reverbatory mechanism as in the attention-switching system).

- A sudden depletion of dopamine leads to an advantage for the indirect pathway that predicts the location of the next number: the cortical representation in the NUMBER map switches to the next number without any corresponding phonological input, like a kind of mental voice.

5. Merging the two Systems

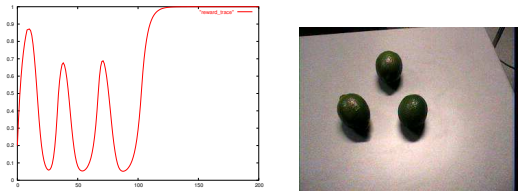

Figure 4: a) Timecourse of the reward signal when three objects are presented in the visual scene. Each depletion corresponds to a switch of focus of attention. b) Corresponding image.

We briefly showed how the sequence-learning system could learn to reproduce a phonological sequence of numbers. After a tonic level of dopamine is applied in SNC to start the counting task, each dopamine depletion switches the cortical representation to the following number. If we add a inhibitory connection from the FOCUS map in the attention-switching mechanism to its REWARD unit, the system focuses each salient point in the image once and then stops. The timecourse of the activity of the REWARD unit is shown in Figure 4 for three objects. It is almost the same timecourse needed for the SNC of the sequence-learning mechanism to reproduce the learned sequence.

Having noticed this analogy, our idea was to link the two systems only by their dopaminergic unit: the REWARD unit of the attention-switching system becomes the SNC unit of the other model and then controls the restitution of the learned sequence. In other words, each time an object is focused, the current number is incremented. At the end, when no more salient object can be found, the sequence-learning

system has stopped on the representation of the exact number of objects in the scene. The cooperation

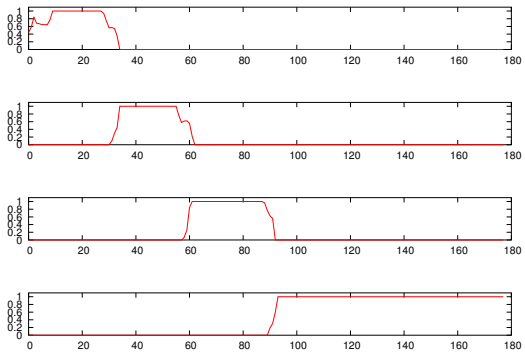

Figure 5: Activity of four neurons in the NUMBER map. From top to bottom, these neurons respond preferentially for the words 'zero' 'one' 'two' and 'three'.

between the two systems has been tested and Figure 5 shows the evolution of the activity of four neurons on the NUMBER map when the reward activity is produced by the attention-switching mechanism. These neurons activate sequentially at each dopamine depletion so that at end of the visual search the only neuron remaining active is the neuron representing "three". This works with one, two or three visual objects, but for more objects we would just need more neurons in the NUMBER map thanks to the distributed architecture of the system.

6. Conclusion

We presented here two computational models that seem in a first view totally independent as they deal with different modalities (vision and reproduction of phonetical sequences), but that can cooperate to produce a new behaviour, namely a cardinal counting task. Our hypothesis was that counting objects in a scene needs to sequentially focus these objects (what is a motor ability) and to associate this sequence with the remembered ordinal sequence of numbers. These two models have two separate basal ganglia channels that communicate via a unique dopaminergic unit, what is coherent with the diffuse innervation of dopamine throughout the Striatum. The first model works in real-time on a robot but for reasons of computational cost the merging of the two systems has only been tested in simulation. Nevertheless, there are some problems: the coding of numbers from phonological inputs is not enough compact and stands for numbers up to ten maybe. We would need another architecture to deal with greater numbers. The sequence-learning system also learns the sequence offline (without the attention mechanism), it would be an interesting feature if the learning occured online.

References

Amari, S.-I. (1977). Dynamical study of formation of cortical maps. *Biological Cybernetics*, 27:77–87.

Berns, G. and Sejnowski, T. (1998). A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience*, 10:108–121.

Brannon, E. M. and Van de Walle, G. A. (2001). The development of ordinal numerical competence in young children. *Cognitive Psychology*, 43:53–81.

Gurney, K., Prescott, T. J., Wickens, J. R., and Redgrave, P. (2004). Computational models of the basal ganglia: from robots to membranes. *Trends in Neurosciences*, 27(8):453–459.

Hikosaka, O., Takikawa, Y., and Kawagoe, R. (2000). Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological Reviews*, 80(3):953–978.

O'Regan, K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24:939–1031.

Piaget, J. (1972). *The psychology of the child*. New York: Basic Books, New York.

Rizzolatti, G. and Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences*, 21(5):188–194.

Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141.

Rougier, N. and Vitay, J. (2005). Emergence of attention within a neural population. *Accepted to Neural Networks*.

Schaal, S., Ijspeert, A. J., and Billard, A. (2003). Computational approaches to motor learning by imitation. *Philosophical Transactions: Biological Sciences (The Royal Society)*, 358(1431):537–554.

Schultz, W., Apicella, P., Scarnati, E., and Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience*, 12:4595–4610.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Science*, 7(7):308–312.

Sugita, Y. and Tani, J. (2002). A connectionist model which unifies the behavioral and the linguistic processes: Results from robot learning experiments. In Stamenov, M. I. and Gallese, V., (Eds.), *Mirror Neurons and the Evolution of Brain and Language*. John Benjamins.

Taylor, J. G. (1999). Neural bubble dynamics in two dimensions: foundations. *Biological Cybernetics*, 80:5167–5174.

Vitay, J., Rougier, N. P., and Alexandre, F. (2005). A distributed model of spatial visual attention. In Wermter, S. and Palm, G., (Eds.), *Neural Learning for Intelligent Robotics*. Springer-Verlag.

Vygotsky, L. (1986). *Thought and language*. MIT Press, Boston.