Attention mechanisms for the imitation of goal-directed action in developmental robots

Ioana D. Goga
Center for Cognitive and Neural Studies
Cluj-Napoca, Romania
ioana@coneural.org

Aude Billard
Learning Algorithms and Systems Laboratory,
Ecole Polytechnique Federale, Lausanne,
Switzerland
aude.billard@epfl.ch

Imitation is a powerful mechanism for social learning and it has received in the last decade, a great deal of interest from researchers in robotics (see Billard and Dillman, 2006). In order to imitate complex behaviors, one must recognize goals, understand how individual actions are embedded in a hierarchy of sub-goals, and recognize recursive structures. Human infants and adults do not copy exactly the movements of the demonstrated act. Deciding *what to imitate* may represent a problem of determining the saliency of objects (Breazeal and Scassellati, 2002), extracting the invariants of the demonstrated acts (Billard et al, 2003) or parsing the structure of the goal hierarchy (Byrne and Russon, 1998).

For instance, when presented with a complex sequence of nesting actions, children aged between 11 and 36 months exhibit different imitation strategies, correlated to their developmental age (Greenfield et al., 1972). During the first stage (12-14 months), infants typically place a single cup in/on a second cup and use a proximity criterion (i.e., same side of the table with the moving hand) for pairing cups. In a second stage (16-24 months) two or more cups are placed in/on another cup and the contiguity criterion is followed (i.e., never reaching behind a nearer cup to use a more distant cup). In the third developmental stage, 28-36-months olds spontaneously imitate using the most advanced nesting strategy, by using a size criterion.

Understanding of the computational mechanisms that underlie the *consistency* of strategic behavior as well as the *variety* of the behaviors occurring within a developmental stage, becomes a crucial step in our quest towards creating autonomous, self-growing cognitive robots. In Goga and Billard (2006) we propose such an account, based on a multiple constraint satisfaction framework, where imitation is addressed as a means for disassembling and reassembling the structure of the observed behavior.

The visual attention system represents an essential component of any cognitive model of goal-directed imitation. Humans selectively direct attention to objects using *bottom-up*, image-based saliency cues and *top-down*, task dependent cues (Itti and Koch, 2001). Objects can gain saliency, due to their properties (i.e., they move quickly, they have bright colors) or due to contextual effects and their behavioral relevance.

During the demonstration phase, attention is employed to break up the visual scene, into a series of smaller chunks that are computationally less demanding. The mechanisms of *joint attention* play a major role in creating a shared context between the demonstrator and the imitator, and in reducing the computational cost of selecting and segmenting possible clues from the environment. Despite an increasing amount of work dealing with joint attention, current research in robotics (Kozima and Yano, 2001; Nagai et al., 2003; Hoffman et al., 2006) concentrates on partial and isolated elements of visual attention behavior, such as simultaneous looking or simple coordinated behavior (for a review see Kaplan 2006). In this work, shared attention ability represents a prerequisite for the development of higher-order cognitive structures required by goal-directed imitation.

The role of the visual model is to integrate bottom-up and top-down constraints in such a way that enables the imitator to follow the demonstrator's focus, to shift attention between different locations in the scene, and to actively select salient objects from the environment. We start with a system capable of simple attention behaviors, such as the selection of objects based on their saliency, gaze following, and skin color detection. In biological systems, *bottom-up attention* is computed in a pre-attentive manner across the entire visual image as a non-linear combination of the contrast with the contextual surround, of different low-level features (Nothdurft, 2000). In this work, two types of feature contrast units (i.e., color and motion) are weighted and then summed into a single *saliency map*.

In principle, *top-down attention* is deliberate and more powerful in orienting attention, and covers the goal-directed factors. Others have shown how a basic set of perceptual preferences (i.e., movement, contrast, color), attention and learning mechanisms are sufficient for gaze following to emerge in typically structured social interactions (Fasel et al., 2002). Hands' gestures are highly informative for the inference of the demonstration goal. The imitator simulated agent has a pre-wired capacity to follow the gaze of the demonstrator and to recognize the skin color of the hand. The model is described in Fig. 1.

An environmental setup for the visual attention model was implemented using the Xanim dynamic simulator (Schaal, 2001). The imitator follows with the gaze the seriation of four differently sized cups, and its task is to reproduce the goal of the demonstration. In Goga and Billard (2006) we describe how a neurobiologically inspired model, developmentally constrained, can account for the systematic differences between infants' strategies.

The weights of the bottom-up and top-down constraints are set to satisfy a number of constraints on the overall shift of attention of the robot. During the learning phase, when information on the demonstrator's direction of gaze signal is available, the weights are adapted in such a way, that: a) *gaze following is preferred* to looking at any static

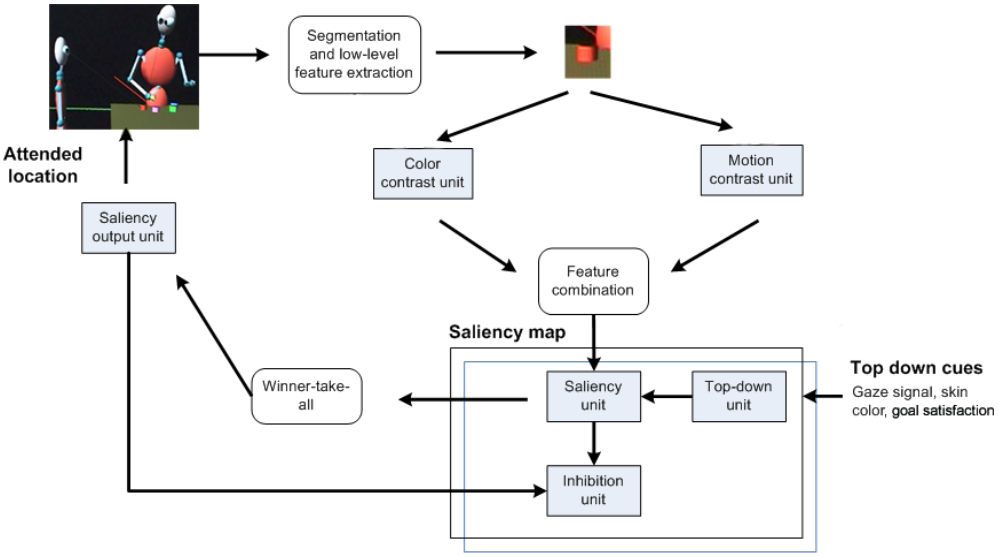

Figure 1. A two-dimensional saliency map is used to deploy the attention to the most salient location in the scene, which is detected using a winner-take-all strategy. The system uses a mechanism of *inhibition of return* to inhibit the attended location and to allow the network to shift to the next most salient object.

object, but not to looking at objects moved by the demonstrator. Accordingly, the imitator closely follows the sequence of movements performed by the demonstrator while it seriates the cups (central and right side in Fig 2). Furthermore, by manipulating the parameters of the attention model, the learner can extract different amounts of information concerning the sequential structure of the demonstrator's behavior.

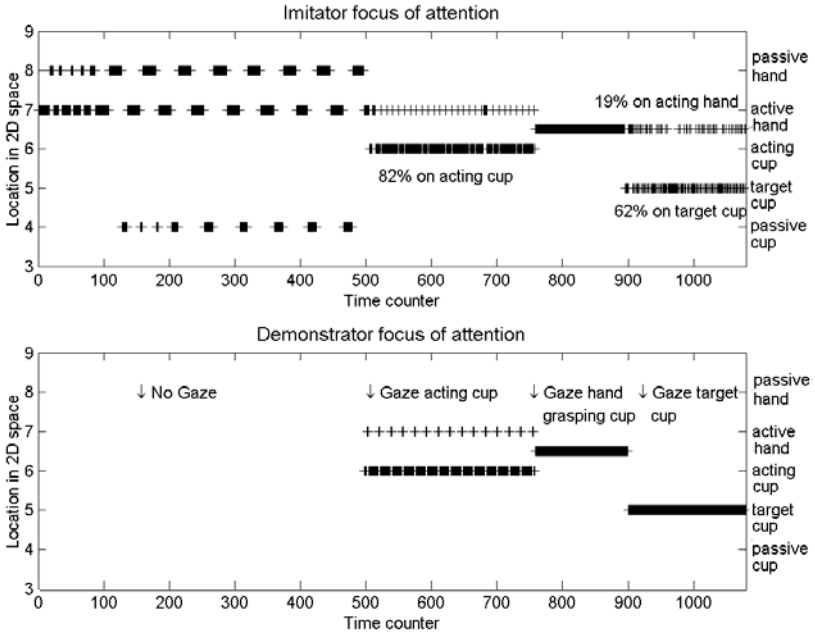

Figure 2. The deployment of the focus of attention for the pair of simulated agents during the demonstration of the seriated cups task. The imitator's focus results from the integration of bottom-up (color, motion) and top-down cues (skin color, demonstrator's gaze).

During imitation, in the absence of the demonstrator's gaze signal, attention is deployed as a result of satisfaction of saliency constraints: b) *skin color preference* (for any static scene, the saliency of an end-effector is higher than that of any colored object); and c) *preference for moving objects* (for any moving object its bottom-up saliency is higher than that of any static object, including the end-effectors) (see the shift of focus between hands and the most salient objects in the left most side of Fig 2). Different imitative behaviors result due to the compound effects of the bottom-up (stimulus-driven) and the top-down (goal-directed) constraints.

Acknowledgements

This work was supported by Swiss National Science Foundation through grant 620-066127 of the SNF Professorships Program. We are very grateful to Stefan Schaal to have provided access to the Xanim simulation environment.

References

A. Billard and R. Dillman (Eds.), 2006, *Special Issue Robotics and Autonomous System,* 54:5.

C. Breazeal and B. Scassellati, (2002), *Imitation in Animals and Artifacts*, MIT Press.

R. Byrne and A. Russon, (1998), *Behavioural and Brain Sciences*, 21, 667-721.

I. Fasel, G. Deak, J. Triesch, and J. Movellan, (2002), *Intl. Conf. on Development and Learning,* MIT.

I. Goga. and A. Billard, (2006), In M. Arbib (Ed.), *Action to Language via the Mirror Neurons System,* Cambridge, MIT Press.

P. Greenfield, K. Nelson, and E. Saltzman, (1972), *Cognitive Psychology,* 3, 291-310.

M. Hoffman, D.Grimes, A.Shon and R. Rao, (2006), *Neural Networks,* Elsevier Science.

F. Kaplan and V. Hafner, (2006), *Interaction Studies, 7.*

H. Kozima and H. Yano, (2001), *Intl. Workshop Epigenetic Robotics,* Sweden.

L. Itti and C. Koch (2001), *Nature Reviews Neuroscience*, 3:2.

Y. Nagai, K. Hosoda and M. Asada, (2003), *Intl. Workshop Epigenetic Robotics, Boston.*

H.C. Nothdurft, (2000), *Vision Research,* 40, 2421-35.

S. Schaal, (2001), *Technical Report*, USC.