An attention framework for learning by interaction in developmental embodied agents

Ioana D. Goga* **Aude Billard****
*Institute for Cognitive and Neural Studies, Coneural
Cluj-Napoca, Romania
ioana@coneural.org

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

Abstract

The role of the visual attention system is to direct computational and behavioral resources toward salient stimuli and to organize behavior around them. This paper presents a biologically inspired model of visual attention mechanisms, based on the integration of multiple bottom-up and top-down constraints. The model has been successfully applied to learning of a complex object manipulation task with a pair of imitator-demonstrator simulated robots.

1 Introduction

Social robotics emphasize the key role of human-robot interaction in building intelligent systems capable to understand and predict human behavior (Billard and Dautenhahn, 2000; Breazeal, 2002). Social abilities, such as imitation, turn taking, joint attention and intended body communication, are fundamental for the development of language and human cognition. In recent years many robotic frameworks have been equipped with imitative skills (Schaal, 1999; Breazeal and Scasselatti, 2000; Kozima and Yano, 2001; Billard, 2002; Demiris and Hayes, 2002). The reason for interest in imitation is obvious: imitative robots offer rapid learning compared to traditional robots requiring laborious pre-programming. *Learning by imitation* requires the capacity to recognize goals, understand how individual actions are embedded in a hierarchy of sub-goals, and extract and recompose recursive structures (Byrne and Russon, 1998). In order to assist the imitator in recognizing the goal of the demonstration and in structuring the imitation behaviour, the teacher can use instruction. In *learning by instruction* the agent is given information about the environment, domain knowledge, or about how to accomplish a particular task on-line.

There is recent interest in robotics, towards collaborative learning based on joint intention, imitation and progressive tutoring of the learner. Breazeal and colleagues (2004) explore multiple forms of socially guided learning, in order to enable robots to quickly learn new skills from human natural instruction and to perform goal-directed tasks in partnership with a human. In similar work, Nicolescu and Mataric (2003) investigate a tutelage-inspired paradigm where a robot learns a sequentially structured task from human demonstration. The human uses short verbal commands to frame the interaction into instruction or demonstration episodes, and provides feedback to correct the robots task model.

We follow this trend by investigating the development in cognitive agents of social abilities using imitation and progressive tutoring. The research described here is part of a long-run project, which explores the means by which a caregiver educates the attention of a learner (human or robot) (Goga and Billard, 2006). Learning by imitation and progressive tutoring can be seen as building blocks of a general bootstrapping mechanism, which uses all available means to focus attention, extract a bit of knowledge, and use this knowledge to perform little more analysis on future inputs. The system starts with a set of pre-programmed behaviors (i.e., gaze following, skin color preference, visuomotor coordination, grasping abilities) and develops in an incremental manner goal-directed behavior and language. The more knowledge the agent acquires through verbal and sensorimotor imitation, the more it can understand instruction and focus its attention. The more it understands, the more it can learn and develop better imitation and cognitive strategies. In this paper, we focus on the description of the attention module, which supports the integration of tutoring and imitation in simulated agents.

2 A general attention framework for interactive learning

An attention framework for collaborative learning should allow the selection of interesting features, the creation

of a shared context between the teacher and the learner, the recognition and categorization of action, and should support the proactive behavior of the system during the imitation phase.

2.1 Joint attention

To benefit from communication and social learning, it is important that both robot and human find the same sorts of perceptual features interesting. Joint attention with a caregiver is one of the abilities that help the infant to direct attention to perceptual structure that makes prominent the relations among objects and caregivers actions, and thus to detect the intention of the caregiver(Tomasello, 1988). In robotics, the investigation of joint attention mechanisms received increased attention during the last decade, due to its crucial role in the development and performance of imitative behaviors. Several researchers attempted to build mechanisms of joint attention inside the robots (Scassellati, 1998; Kozima and Yano, 2001; Breazeal and Scassellati, 2000; Nagai et al., 2003). Robotic systems, such as those of Scasellati (1998) and Demiris and Hayes (2002) are able to track the gaze of a human instructor and to imitate the motion of the instructors head.

The ability of joint attention develops gradually. Researchers in cognitive sciences and robotics agree that the main steps in producing the mechanisms of joint attention are: a) recognition and maintenance of *eye contact*; b) engagement in joint attention through *gaze following*; c) *imperative pointing* used to obtain an object that is out of reach by pointing at that object; and d) *declarative pointing* used to draw attention to a distal object (Scasselatti, 1998). Nagai et al. (2003) illustrated how based solely on visual attention mechanisms and learning with self-evaluation, a robot can acquire sensorimotor coordination through a staged developmental process similar to the developmental path that human infants undergo.

We take into account the results of previous work in the modeling of joint attention behavior and we focus on the investigation of the role that joint attention mechanisms may play in the extraction and reproduction of the imitation goals. Recently Kaplan and Hafner (2006) emphasized the importance of moving towards the higher-order, cognitive aspects of joint attention behavior. Joint attention is more than gaze following or simultaneous looking attention, and it can be better defined as *'a coordinated and collaborative coupling between intentional agents where the goal of each agent is to attend to the same aspect of the environment'*. For an agent to learn from social interaction, it must be able to detect and to manipulate the focus of attention of other agents, and to engage in coordinated interaction.

We use simulated robots acting in a controlled environment, and we implement the prerequisites for joint attention behavior. The demonstrator can be programmed to execute different types of tasks (translation movements, push or hit objects, seriate a number of cups, describe objects verbally; see also Billard et al., 2003). The imitator is able to follow the gaze of the demonstrator and to detect the movements of the hands. Detection of the hands provides valuable information on the course of the action currently taken by the demonstrator (i.e., whether the hand approaches an object to grasp it, it manipulates the object, or it approaches a target while grasping the object). Sharing the visual context with the demonstrator helps the imitator to infer the goal of the action. A general constraints satisfaction framework is implemented to account for different attention behaviors: select objects from environment and keep the focus, follow the caregiver's gaze and detect the movements of the end-effectors, perform shifts of attention between different locations of interest.

2.2 Attention for recognition and learning

During demonstration the learner must recognize the movements performed by the demonstrator, infer the goal of its behavior and learn an internal model for further reproduction. The attention system cannot be investigated separately from processes such as object recognition and categorization. One of the classical models dealing with attention for categorization is the Adaptive and Resonance Network ART (Carpenter and Grossberg, 1998). A central feature of all ART systems is a pattern matching process that compares an external input with the internal memory of an active code. ART matching leads either to a *resonant* state, which persists long enough to permit learning, or to a parallel memory search. If the search ends at an established code, the memory representation may either remain the same or incorporate new information from matched portions of the current input. If the search ends at a new code, the memory representation learns the current input.

More recently, Marsland et al. (2005) developed a Growing-When-Required GWR neural network, to deal with novelty detection for on-line learning in mobile robots. A neural network inspects each new input and evaluates it for novelty with respect to data already perceived. For each node a measure of habituation is maintained, which gives an indication of the familiarity with that stimuli. When fully habituated, nodes ignore further stimulation and do not get updated. When the orienting response is reinstated either due to novelty detection, when a new node is created, or due to forgetting, when a node is dishabituated, the information goes through the focus of attention, to higher level processing, such as learning. Maistros and colleagues (2001) integrated the GWR network with a schema network that handles the perception-action coupling for an imitation task. The attention system is used to cluster the prin-

cipal components of the demonstrated behavior and to activate the corresponding nodes in the network for the recall.

In Goga and Billard (2006) a layered neural architecture is implemented for categorization and sequence learning. The sensorial external maps operate on a short time scale, in the sense that they respond spontaneously to external inputs. At this level of visual awareness, the systems capacity to store associations is limited to the duration of the short-term memory. A saliency signal received by an object which is in the focus of attention enhances the object's neural representation and enables the creation of a new node, in a categorical, higher-level processing layer. A *vigilance parameter* inspired by the adaptive resonance theory is used to weight how close an input must be to the prototype for matching to occur, and when a new representation node should be created. The categorical layer operates on a larger time scale, allowing the system to extract and store temporal sequences with various time lags. During the demonstration, statistical regularities about objects located in the focus of attention are extracted in different sets of weights.

2.3 Attention for goal-directed imitation

During imitation, the learner composes the actions extracted in the internal model in order to reproduce a part or the entire sequence demonstrated by the teacher. 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, 2000), 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.

We simulated the imitation of the seriated cups task with a pair of humanoid robots. The demonstrator seriates a set of 5 cups, by using a sub-assembly strategy (i.e., a previously constructed structure consisting of two or more cups is moved as a unit in another cup or cup structure). The imitator follows the demonstration and its task is to reproduce the sequence of actions. During the imitation phase, the attention system parses the objects in the environment and computes their feature-based saliency (i.e., color contrast or motion contrast). Objects which enter the focus of attention can activate the corresponding nodes in the categorical layer, and eventually one of the goals of the system is activated through bottom-up stimulation. The goal sets the type of the action (i.e., grasp or move), which is further executed by the system through a process of successive operations aimed at minimizing the distance between the current state of the world and the desired state corresponding to the goal. As a result of probabilistic satisfaction of multiple constraints, the imitator is able to reproduce a variety of imitative behaviors, in a similar manner with the human infant.

3 The visual attention model

Our approach to computational modeling of visual attention draws inspiration from different sources. Investigations in the visual system processes suggest that the control inputs to the attention mechanism can be divided into two categories: stimulus-driven (or bottom-up) and goal-directed (or top down) (Itti and Koch, 2001).

3.1 Bottom-up attention

Bottom-up attention is computed in a pre-attentive manner across the entire visual image. The bases of bottom-up computational models are the experimental results obtained using the Feature Integration theory of Treisman and colleagues (Treisman and Gelade, 1980). The first neurally plausible computational architecture for controlling visual attention was proposed by Koch and Ullman (1985), whose model was centered around a *saliency map* concept. The map calculates saliency, that is, stimulus conspicuity, at every location in the visual scene, based on low-level features of the object. A *winner-take-all* approach is then used to decide on the most salient part of the scene (Koch and Ullman, 1985; Khadhouri and Demiris, 2005).

We implemented a two-component framework consisting of a saliency map that controls the deployment of attention on the basis of bottom-up saliency and top-down cues. The focus of attention is deployed to the most salient location in the scene, which is detected using a winner-take-all strategy (See Figure 1). Once the most salient location is focused, 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 (Itti and Koch, 2001). Computationally, inhibition of return implements a short-term memory of the previously visited locations and allows the attentional selection mechanism to focus instead on new locations. An alternative mechanism which prevents attention from

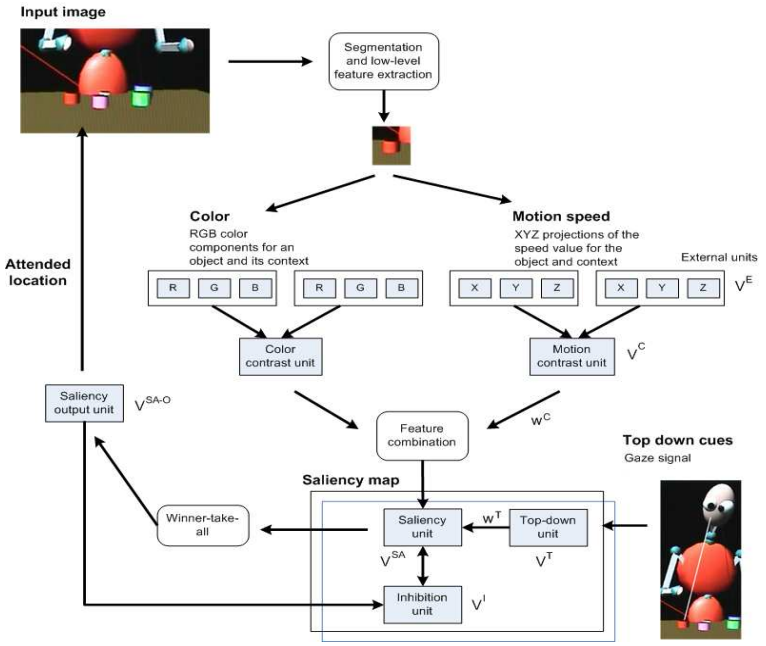


Figure 1: Model of visual attention based on the integration of bottom-up saliency and top-down constraints. Low-level features (color and motion) are extracted from the visual image for each object and its context. Feature contrast is combined in a two-dimensional saliency map for all the objects in the environment and for the end-effectors. Top-down cues and inhibition of return modulate the activity of the saliency map. Attention is deployed to the most salient location, selected through a winner-take-all mechanism.

permanently focusing on the most active location is that of habituation implemented by Marsland et al. (2005) and Maistros et al. (2001).

Some of the most common pre-attentive features suggested by researchers in their theories are color, orientation, luminance, depth and motion. What seems to matter in driving the bottom-up attention is the contrast of the features with respect to the contextual surround, rather then the absolute values of the features (Nothdurft, 2000). In our model, saliency is computed based on the linear integration of contrast of color and motion.

Each contrast unit V_i^{C} receives input from two pairs of external units V_j^{E}, that encode color components (R,G,B) or the projections on (X,Y,Z) axes of the object's speed of motion. Color contrast is computed using the value of the sensor $n_{j,l}$ at the object location l and $N = 6$ values corresponding to 6 contact points with the surrounding context. Motion contrast is computed using one reading of the speed value for the object and N context readings $n_{j,l_{context}}$ corresponding to the speeds of all objects in the visual image. The output of an external unit V_j^{E} corresponding to the component j read at location l is given by:

$$V_j^{\mathrm{E}}(t,l) = \begin{cases} n_{j,l}, & \text{if location l is visible at t,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

$$V_j^{\mathrm{E}}(t, l_{context}) = \frac{1}{N}\sum_{k=1}^{N} n^k_{j,l_{context}}. \tag{2}$$

The output of a contrast unit V_i^{C} is given by the euclidean distance between the components of the feature corresponding to the object location l and to the context surrounding the object:

$$V_i^{\mathrm{C}}(t,l) = \mathcal{F}\Big(\sqrt{\sum_{j=1}^{3}(V_j^{\mathrm{E}}(t,l) - V_j^{\mathrm{E}}(t, l_{context}))^2}\Big) \tag{3}$$

where $\mathcal{F}$ is the sigmoid output function $\mathcal{F}(x) = 1/(1+e^{-x})$ used for all units.

The output of a saliency unit V_i^{SA} is given by the weighted summation of the saliency features $j \in C$ and the contribution of the top-down cues $k \in T$:

$$V_i^{\mathrm{SA}}(t,l) = \mathcal{F}\Big(\sum_{j=1}^{3} w_j^{\mathrm{C}} \cdot V_j^{\mathrm{C}}(t,l) + \sum_{k=1}^{1} w_k^{\mathrm{T}} \cdot V_k^{\mathrm{T}}(t,l) - V_i^{\mathrm{I}}(t,l)\Big) \tag{4}$$

The output of a top-down unit is $V_k^{\mathrm{T}}(t,l) = 1$ if the location l is referred by the top-down cue k and 0 otherwise. The weights w^{C} stand for the gains of object's color, speed of object's motion and skin color. The top-down w^{T} weight stands for gaze following. The weights

are independent of i (object identity) and result from the satisfaction of the attention constraints (see Section 3.2).

Neurons in the saliency map compete according to a winners-take-all strategy and the winning unit i sets the activity of its output unit to 1:

$$V_i^{\text{SA-O}}(t,l) = \begin{cases} 1, \text{ if } V_i^{\text{SA}}(t,l) > V_j^{\text{SA}}(t,l), \forall j \neq i \\ 0, \text{ otherwise.} \end{cases} \quad (5)$$

The activity of the inhibitory unit is a function of the input received from the salient unit V_i^{SA} if it corresponds to the winner location, otherwise its activity is a decayed memory of its previous activation:

$$V_i^{\text{I}}(t,l) = \begin{cases} \left(\exp(\frac{t}{f(V_i^{\text{SA}}(t,l))} - a) - b\right) \cdot V_i^{\text{SA}}(t,l), \\ \quad \text{if } V_i^{\text{SA-O}}(t,l) = 1 \\ \tau_i \cdot V_i^{\text{I}}(t-1,l), \text{ otherwise,} \end{cases} \quad (6)$$

where τ_i is the time decay rate of inhibitory unit i. Parameters a and b are set so that inhibition increases from 0 to a maximum value equal to V_i^{SA}, when it shunts down the salient unit. The larger the value of the saliency winner unit, the longer it will stay active, but also the higher will be its inhibition. After shutting down the salient unit, the inhibitory unit preserves a memory of its activation, which decays in time and allows the unit to win again further in future.

3.2 Integration with top-down constraints

Top-down attention is deliberate and more powerful in directing the attention. Wolfe (Wolfe and Gancarz, 1996) constructed a flexible model of human visual search behavior that uses a top-down mechanism to control bottom-up features, which are relevant to the current task. Visual stimuli are filtered by broadly-tuned channels (such as color and orientation) to produce *feature maps* with activation based upon both bottom-up and top-down demands. The feature maps are combined by a weighted sum to produce an activation map.

Using a similar concept, Breazeal (2002) has augmented a vision system (described in Scassellati, 1998) with facial features for emotive expression. The implementation focuses on three pre-attentive processes: color, motion, and face pop-outs represented in bottom-up feature maps, which are further combined with a habituation function to produce an attention activation map. Top-down influences from motivational and behavioral sources, combine with bottom-up habituation effects to bias the robot's gaze preference. For instance, when the top-down social drive is activated by face stimuli, the face gain is influenced by the *seek people* and *avoid people* behaviors. This result in a system that directs eye gaze based on current task demands.

In our model, during the demonstration phase the top-down constraints support the formation of a shared attention context between the teacher and the learner. The focus of attention is deployed as a function of the satisfaction of the bottom-up and top-down constraints and of the functioning of the inhibition mechanisms. In the lack of top-down cues, attention is deployed as a result of satisfaction of bottom-up saliency constraints (i.e., preference for moving objects and for the skin color). The model can be easily extended to integrate other types of constraints.

- **Skin color preference**. For any static scene, the bottom-up saliency of the hand is higher than that of any object.
- **Preference for moving stimuli**. For any moving object, its bottom-up saliency is higher than that of any static object, including the hands.
- **Motion versus skin color preference**. Saliency of a moving object is smaller than the saliency of a hand moving at comparable speed.
- **Gaze following versus moving objects**. The global saliency of any static object located in the focus of attention is higher than the bottom-up saliency of any moving object located outside the focus.
- **Gaze following versus moving hand**. The bottom-up saliency of a moving end-effector is higher than the saliency of a static object located in the focus of attention.

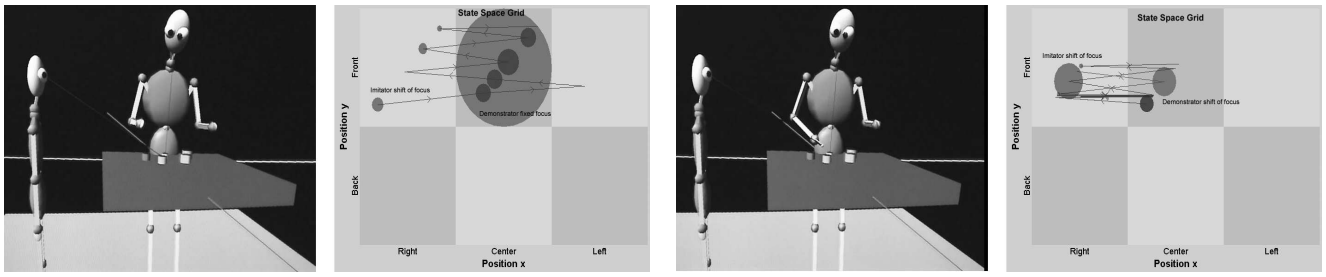
Based on these constraints, we compute the weights w^{C} for object's color, speed of object's motion and skin color, and the weight w^{T} for gaze following.

4 The simulation environment

An environmental setup for the joint attention model was implemented using Xanim dynamic simulator (Schaal, 2001), to model a pair of 30 degrees of freedom (Head 3, Arms 7 *2, Trunk 3, Legs 3*2, Eyes 4 D.O.F.) humanoid robots. The simulated robot is controlled from Cartesian states through inverse dynamics and inverse kinematics servos. The external force applied to each joint is gravity. There is no collision avoidance module. The environment is controlled, in other words, only a predefined set of objects and end-effectors are visually recognized and manipulated. A motor servo is used to read the current state of the robot/simulation (i.e., position, color, orientation and rotation angles, and motion speed) and to send commands to the robot/simulation.

5 Attentional behavior

The attention system was tested during the simulation of the seriation cups task with the pair of demonstrator-imitator agents. The deployment of the focus of atten-



(a) (b)

Figure 2: Shifts of focus of attention for the pair of demonstrator-imitator simulated robots, illustrating the functioning of the multiple constraints framework. A state space grid is used to display the duration values for each event in the appropriate location cell. The connectivity arrows show the flow of real time. a) The attention of the demonstrator is focused on the cup that it intends to grasp (in the center of the grid). The imitator's focus of attention shifts between the acting cup and the acting hand of the demonstrator. b) While approaching the acting cup, the attention of the demonstrator shifts between the acting hand and the position of the cup, and the imitator closely follows it.

tion of the demonstrator is a good predictor of its intentions. During the seriation of two objects, the teacher gazes successively, the acting hand, the acting cup, the hand manipulating the cup and the target cup. By following the demonstrators gaze, the imitator learns how to segment the sequence of behaviors, which are the predecessors of each action, and how to recompose the sequence of movements during imitation.

If the demonstrators gaze signal is not available, attention is deployed as a result of the satisfaction of saliency constraints. For any static scene, the saliency of an end-effector is higher than that of any colored object, and the focus of attention of the imitator shifts between the locations of the demonstrator's hands. In other words, in the absence of top-down cues from the demonstrator (i.e., gaze or speech), its hand actions carry the most valuable information, which are available to the imitator to infer its intentions.

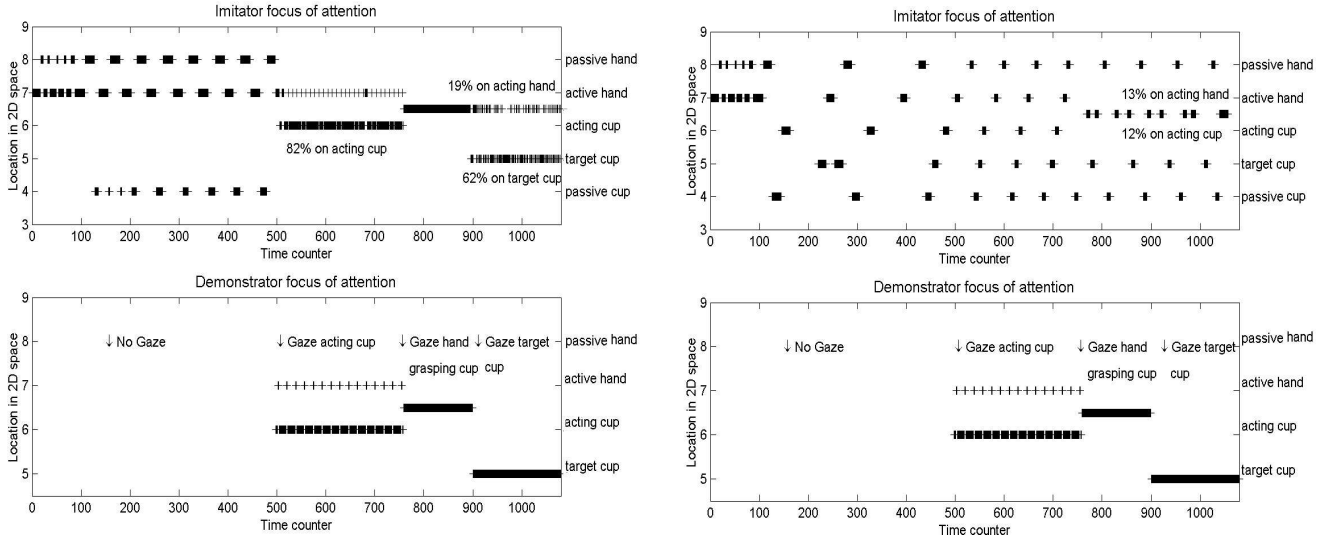
When the demonstrators gaze signal is available, the weights are adapted in such a way, that *gaze following* is preferred to looking at any static object. This behavior is illustrated in Figure 2a. The imitator learns about the affordances of an object (i.e., an object can be grasped by an empty hand) as well as the effectivities of its hand (i.e., how to grasp it). When the hand approaches the object, the *preference for moving objects* equals the effects of gaze following and the imitator's focus of attention shifts between the moving hand and the location of the acting cup (Figure 2b). By paying attention to the hand's movements, the imitator learns how to shape the hand (i.e., rotate the end-effector and lift the object from bellow) in order to grasp the object.

We were interested in comparing the effects of modulating the weights of the constraints on the imitative behavior of the learner. In Figure 3 is depicted the timing between the demonstrator's attention behavior and the imitator's shift of focus. When the gaze signal is not available (left side of Figure 3), the attention of the imitator can shift between the hands and one or several colored objects in the environment. By increasing the gain of the color constraint and the inhibition of return, all the objects posted on the table can be gazed alternatively. During imitation, this is a desired attentional behavior, which allows the robot to parse all objects before activating its internal goal and choosing the acting and target cups.

When the gaze signal is available, the imitators focus of attentions closely follows the gaze of the demonstrator (central and right side in Figure 3a). The attention mechanism helps the learner to extract the sequential structure existent in the demonstrated actions (i.e., grasp the acting cup and move it to the location of the target cup). The amount of time spent in gazing each object or segment of behavior determines the strength of the precedence links created. An empty hand and a colored object activate the motor schema for grasping the object, which in turn activates a translation motion of the hand holding the cup, towards the location of the target cup. If the weights of all constraints in the system are decreased, the imitator focuses equally on all objects, without being able to select the relevant stimuli. This behavior is illustrated in Figure 3b.

6 Discussion and future work

The current model can be enhanced by addressing the instability of the attentional focus. This occurs due to the operation on a low-time scale of the mechanism of inhibition of return. One possibility is to implement a mechanism that learns to predict which will be the



(a) (b)

Figure 3: Overall times spent in gazing different stimuli from the environment by the demonstrator-imitator pair of simulated robots. a) The imitator closely follows the focus of attention of the demonstrator, when this is available. Time percents are indicated for different phases of the demonstration. b) All gains in the system are decreased, and the imitator focus of attention shifts from one object to another.

shared focus of attention in the next moments and thus, to maintain the joint focus of attention over larger time periods. Another improvement would consist in allowing the imitator to continuously adapt the weights of the attention constraints. In different phases of the task, the imitator can learn to weight differently the top-down and the bottom-up constraints, as a function of the adaptive value carried out by the demonstrator's actions.

In future work the attention system presented here will be developed to support progressive tutoring of the learner. Caregivers assist their infants to perceive referring actions by providing them timely feedback and guidance. Besides pointing, researchers in developmental psychology described the usage of different gestures that direct the childs attention, narrow the search space and enhance the speed of achieving a common understanding. According to Zukow-Goldring (2003), five gestures that direct attention often accompany caregivers' verbal messages: *embody*, *show*, *demonstrate*, *point*, and *look*. Longitudinal data suggest that caregivers of less advanced infants (not necessarily younger infants) use *embody* and *show* most frequently, shifting to *demonstrations*, *points*, and eventually *looks* as the infant develops (Zukow-Goldring, 1997). There is large, unexploited learning potential in the usage by humans of the embody gesture to teach the robot the dynamics of a movement.

We are currently investigating the strategies used by human caregivers to scaffold the experience of their infants during the execution of a complex task in collaboration. In order to implement these processes in robots, the demonstrator robot should be able to continuously tune its verbal, motor and attention behavior to the reactions of the learner. On the other hand, if we are to build an infant robot capable to work in collaboration with a human, we have to enable it to provide feedback. We intend to develop a strategic attention behavior for the demonstrator, which will allow it to follow and to achieve its goals, as well as to detect and to respond in a timely fashion to the feedback provided by the learner. The learner will help the instructor by expressing its internal state via communicative acts (i.e., speech, hand's gestures). The ecology, and in the same time the novelty of our approach results from the investigation of these issues based on real data transcripts, which characterize the sensorimotor and linguistic patterns of interaction between human caregivers and infants aged between 1 and 3 years (Goga and Billard, 2006).

7 Acknowledgments

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

8 References

Billard, A., Epars, Y., Schaal, S., Cheng, G. (2003). Discovering Imitation Strategies through Categorization of Multi-Dimensional Data, *Procs. Int. Conference on Intelligent Robots and Systems IROS 03.*

Billard, A. (2002). Imitation: a means to enhance learning of a synthetic proto-language in an autonomous robot, in *Imitation in Animals and Artifacts*, (K. Dautenhahn and C. L. Nehaniv , Eds.), MIT Press, pp. 281-311.

Billard, A., Dautenhahn, K. (2000). Experiments in social robotics: grounding and use of communication in autonomous agents, *Adaptive Behavior*, vol. 7, 3/4.

Breazeal, C., Hoffman, G., Lockerd, A. (2004). Teaching and Working with Robots as a Collaboration, in *AAMAS 2004.*

Breazeal, C. (2002) *Designing social robots.* Bradford book MIT Press, Cambridge, MA.

Breazeal, C., Scassellati, B. (2000). Infant-like social interactions between a robot and a human caregiver, *Adaptive Behavior*, 8(1), pp. 49-74.

Byrne, R. W., Russon, A. (1998). Learning by imitation: A hierarchical approach. *Behavioural and Brain Sciences*, vol. 21, pp. 667-721.

Carpenter, G.A., Grossberg, S. (1998). Adaptive Resonance Theory, in M. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*, Second Edition, MIT Press.

Demiris, J., Hayes, G.M. (2002). Imitation as a dual-route process featuring predictive and learning components: A biologically plausible computational model, in *Imitation in Animals and Artifacts* (K. Dautenhahn and C.L. Nehaniv, Eds.), pp. 321-361, Cambridge, MA: MIT Press.

Goga, I., Billard, A. (2006). Development of goald directed imitation, object manipulation skill and language in humans and robots, In M. Arbib. (Ed.) *Action to Language via the Mirror Neuron System*, Cambridge, MA.

Greenfield, P., Nelson, K., Saltzman, E. (1972). The development of rulebound strategies for manipulating seriated cups: a parallel between action and grammar, *Cognitive psychology*, 3:291310.

Kaplan, F., Hafner, V. (2006). The challenges of joint attention.*Interaction Studies*, 7(2).

Khadhouri, B., Demiris, Y. (2005). Compound effects of Top-down and Bottom-up influences on Visual Attention during Action Recognition, in *Proceedings of IJCAI-2005*, pp. 1458-1463, Edinburgh.

Koch, C., Ullman., S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4, pp. 219.227.

Kozima, H., Yano, H. (2001). A robot that learns to communicate with humancaregivers, *Procs of the First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, Lund, Cognitive Studies 85.

Itti, L., Koch, C. (2001). Computational modeling of visual attention, *Nat. Rev. Neuroscience*, 2(3).

Maistros, G., Marom, Y., Hayes, G. (2001). Perception-Action Coupling via Imitation and Attention. In *Proceedings of the AAAI Fall Symposium on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, Cape Cod, MA.

Marsland, S., Nehmzov H., Shapiro, J. (2005). Online Novelty Detection for Autonomous Mobile Robots, *J. Robotics and Autonomous Systems*, Vol 51, pp. 191-206.

Nagai, Y., Hosoda, K., Asada, M. (2003). How does an infant acquire the ability of joint attention?: A constructive approach, *3rd Int. Workshop on Epigenetic Robotics (EpiRob '03)*, pp. 91-98.

Nicolescu, M., Mataric, M. (2003). Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *Procs of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, Melbourne, Australia.

Nothdurft, H.C. (2000). Salience from feature contrast: additivity across dimensions, *Vision Research*, 40.

Scassellati, B. (1998). Imitation and mechanisms of joint attention: a developmental structure for building social skills on a humanoid robot, *Autonomous Agents Workshop 98.*

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Science*, 3, pp. 223:231.

Schaal, S. (2001). The SL simulation and real-time control software package, Technical Report Computer Science Tech Report, University of Southern California.

Tomasello, M. (1988). The role of joint attentional processes in early language development, *Language Sciences*, 1.

Treisman, A.M., Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology* 12, pp. 97-136.

Wolfe, J.M., Gancarz, G. (1996). Guided search. In *Basic and Clinical Applications of Vision Science*, pp. 189.192. Kluwer Academic Publishers.