Integrating Reinforcement-Learning, Accumulator Models, and Motor-Primitives to Study Action Selection and Reaching in Monkeys

Dimitri Ognibene (dimitri.ognibene@istc.cnr.it)
Francesco Mannella (francesco.mannella@istc.cnr.it)
Giovanni Pezzulo (giovanni.pezzulo@istc.cnr.it)
Gianluca Baldassarre (gianluca.baldassarre@istc.cnr.it)
Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche,
Via San Martino Della Battaglia 44, 00185 Roma, Italy

Abstract

This paper presents a model of brain systems underlying reaching in monkeys based on the idea that complex behaviors are built on the basis of a repertoire of motor primitives organized around specific goals (in this case, arm's postures). The architecture of the system is based on an actor-critic reinforcement-learning model, enhanced with an accumulator model for action selection, capable of selecting sensorimotor primitives so as to accomplish a discrimination reaching task that has been used in physiological studies of monkeys' premotor cortex. The results show that the proposed architecture is a first important step towards the construction of a biologically plausible integrated motor-primitive based model of the hierarchical organization of mammals' sensorimotor systems.

Introduction

This paper aims to present a model based on the hypothesis that sensorimotor systems of organisms are organized on the basis of repertoires of sensorimotor primitives (Arbib, 1981) that are suitably "assembled" to produce complex behaviors. This hypothesis is important both for understanding organisms' functioning and for building artificial intelligent systems (Schaal, 1999; Rainer & Tani, 2004).

The first specific goal that the model pursues is to understand how organisms build repertoires of motor primitives on the basis of experience. In particular, the paper focuses on the acquisition of motor primitives related to the production of arm's postures in space. With this regards, it has been shown that neural systems of various animal species, from insects and amphibians to mammals and humans, are organized around motor primitives that, when triggered, tend to accomplish a particular *goal*. For example Giszter, Mussa-Ivaldi & Bizzi (1993) showed that if some regions of the spinal cord of frogs are stimulated electrically, their lower limbs tend to perform movements so as to achieve a specific resting point (posture) independently of the starting position. Remarkably, there seem to be a relatively small number of these motor primitives, whose origin is likely filogenetic, encoded in the spinal cord. Similarly, Graziano, Taylor & Moore (2002) showed that if the premotor cortex of monkeys is stimulated electrically, their arms tend to assume a given posture in space. In humans a great part of low-level sensorimotor skills are acquired during the first years of life without direct rewards and on the basis of self-generated experience (von Hofsten, 1982). These skills involve the capability of both assuming postures in space and generating cyclical movements (through "central pattern generators", cf. Swanson, 2005, and Schaal, 1999: the latter will not be tackled here).

A second specific goal of the paper is to study how organisms can assemble motor primitives to accomplish complex tasks. Increasing evidence is showing that basal ganglia (Kandel, Schwartz & Jessell, 2000) might play an important role in this process (Nakahara, Doya & Hikosaka, 2001; Baldassarre, 2002). Basal ganglia are nuclei that form the basis of vertebrates' forebrain. They receive signals from virtually the whole cortex and send signals, via the thalamus, to the motor part of it (pre-frontal, premotor and motor cortex). Basal ganglia's dopaminergic neurons are involved in classical conditioning tasks where, by experience, originally neutral stimuli progressively acquire the role of predictors of *primary rewards* (Shultz, Dayan & Montague, 1997). These processes have been successfully modeled on the basis of *actor-critic* reinforcement-learning architectures (Barto, Sutton & Anderson, 1983). In particular, the functioning of the "critic", based on the TD-learning algorithm (Barto & Sutton, 1998), has been shown to mimic some aspects of the physiology of basal ganglia's dopaminergic neurons (Houk, Davis & Beiser, 1995).

The third goal of the model is to start to develop an actor that is both more soundly related to the brain's functioning and closely integrated with the motor-primitive system. The reason is that the "actor" component of the actor-critic architecture, that should mimic the basal ganglia's sensorimotor function, has been modeled and related to known brain's physiology in much less detail with respect to the critic (Joel, Niv & Ruppin, 2002). The way that will be followed to pursue this goal is suggested by Schall (2001) who studied monkeys that accomplish oculomotor saccades to one of few alternative targets. In these monkeys some neurons of the frontal-eye field (premotor cortex) give place to a *race* in which different (groups of) cells "accumulate evidence" (activate) in favor of the different options: the first (group of) cell(s) that reaches a given threshold triggers a saccade towards the corresponding target. These processes have been modeled through *accumulator models* (e.g., cf. Usher & McClelland, 2001), probably the best available biologically-plausible models of action selection and reaction times. The actor presented here assumes that basal ganglia fuel a race in the premotor cortex and in this way they select the motor primitives to execute.

The architecture presented here has been trained and tested using the "discrimination reaching task" used by Cisek & Kalaska (2005) to carry out physiological recordings in monkeys' premotor cortex (this task showed to trigger "races" among the premotor neurons similar to those mentioned above). The task is composed of five phases (see bottom part of Figure 5): (1) *center-hold time – cht*: the monkey's hand is positioned on a manipolandum at a central starting position of an horizontal plane, and a green cue circle appears at the center of a screen set in front of the subject; (2) *spatial cue – sc*: a red and a blue circle (with a 2 cm radius) appear on the screen at two opposite positions of eight possible target locations distributed around a circle; (3) *memory – mem*: a green cue circle appears again at the center of the screen; (4) *color cue – cc*: a color cue, either red or blue, appears at the center of the screen: this non-spatial cue signals which of the two memorized color-coded spatial cue locations is the target that the monkey should reach; (5) *go signal – go*: eight green circles appear at all the possible target locations: if the monkey reaches the target position that matches both one of the two spatial cues *and* the color cue, it receives a reward. In the simulations, the first four phases last 1 s. each, while the fifth lasts 16 s.

The rest of the paper first presents the architecture of the model, then presents the results obtained with it, and finally concludes illustrating the model's strengths and weaknesses.

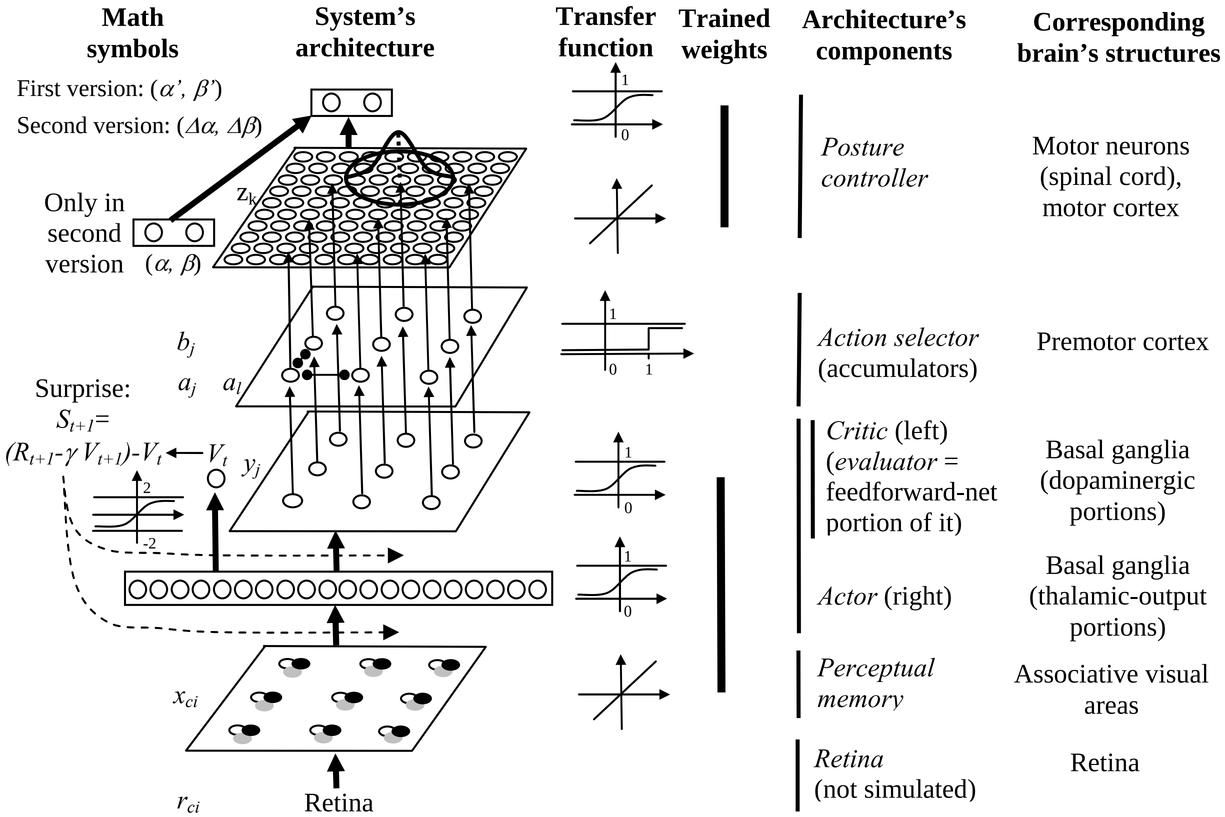

Figure 1: Architecture's components and corresponding brain areas. Symbols: empty, gray and black small circles (in perc. mem.): cells sensitive to green, red and blue; bold arrows: all-to-all connections; arrows: one-to-one connections (weight = +1); dot-head arrows: reciprocal inhibition connections (only few of them have been drawn); dotted arrows: surprise learning signal. The Gaussian function in the posture controller's map is an example of one cell's receptive field.

Methods

The body of the system is made up by a two-segment arm, that moves on a 2D plane (Figure 2), and a retina (see below). The two segments of the arm measure 20 cm each. Each segment has one degree of freedom: the upper arm can move 180° with respect to the system's torso by pivoting on the shoulder joint, while the forearm can move 180° with respect to the upper arm by pivoting on the elbow joint (only simple kinematics of the arm, and no dynamics, were simulated).

The controller of the system (Figure 1) is formed by four main components: perceptual memory, actor-critic, action selector, and posture controller. The "life" of the system has two phases: (1) *childhood phase*: the system learns to produce suitable movements to bring the arm to desired postures on the basis of self-generated experience: this phase updates the posture controller's weights; (2)

adulthood phase: the system learns, by reinforcement learning, to accomplish the discrimination-reaching task: this learning phase updates the actor-critic's weights. Now the architecture and functioning of the system's components and the two learning phases are explained in detail.

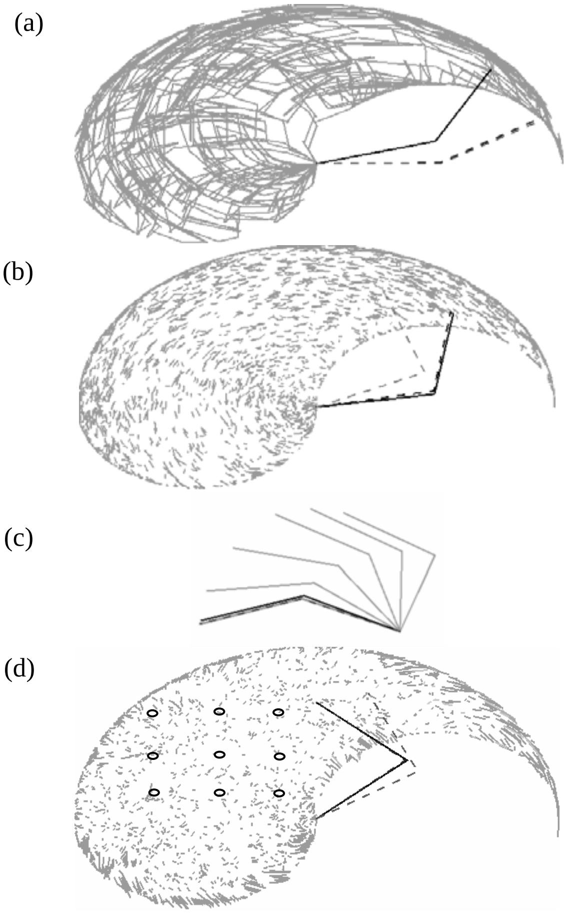

Figure 2: Errors (gray short segments) of the posture controller's hand between desired postures (black arms) and actual postures (dark dashed arms), achieved from various initial postures (light dashed arms): (a) posture controller's first version: before training,; (b) first version: after training; (c) first version: during a generalization test where the controller reaches a target posture in more than one step; (d) second version: after training; dots represent the position of the 9 units used in the discrimination reaching task.

Perceptual memory This is a map of 27 units x_{ci} ($c \in \{g, r, b\}$; $i \in I$, $|I|$=9) assumed to correspond to visual cortex areas. These units receive topological excitatory connections from the retina (not simulated here) and respond to stimuli with 3 different colors (green, red, and blue) and 9 different positions on the retina. These units are "leaky-integrators":

$$x_{cit+1} = \min\left[\left(x_{cit} + \tau\left(-\kappa x_{cit} + r_{cit+1}\right)\right), 1\right]$$

where min[.] keeps $x_{ci} < 1$, τ is an integration time step (τ = 0.1), κ is a decay coefficient (κ = 0.1), and r_{ci} is a retina's signal caused by the on-off colored cues ($r_{ci} \in \{0, 10\}$).

Actor-critic This is a standard feed-forward network with 25 hidden units, 27 input units (x_{ci}), and 10 output units. Nine output units (*actor*), located on a 2D map and with a Sigmoid transfer function, select one of the 3×3 possible spatial targets for the arm (see introduction). The last output unit (*evaluator* part of the *critic*), with a linear transfer function, produces evaluations V_t of perceived states (see below). The actor-critic is a neural implementation of the actor-critic architecture (Sutton and Barto, 1998). The evaluator's output is used to compute the critic's surprise S_{t+1}, used to train both the actor and the evaluator (see below), on the basis of the reward R_{t+1} and evaluations produced at couples of succeeding states:

$$S_{t+1} = \left(R_{t+1} + \gamma V_{t+1}\right) - V_t$$

where γ is a discount factor (γ = 0.99).

Action selector The action selector is composed of a 2D map of *accumulator* units with activation b_j and activation potential a_j ($j,k \in A$, $|A|$ = 9: for simplicity, in this research only 9 units corresponding to the 8 targets, plus the central starting position, were used). Each of these units is activated by one topologically corresponding unit of the actor. The accumulator units have lateral inhibitions and give place to a (noisy) competition that integrates in time, and amplifies the differences, of the signals coming from the actor. The unit whose a_j reaches a threshold T (T = 1) wins the competition, activates with b_j = 1 (the units have a step activation function), and triggers the pursuing of the goal-posture in the posture controller:

$$a_{jt+1} = \max\left[\left(a_{jt} + \tau\left(-\delta\, a_{jt} - \iota \sum_{k \in A, k \neq j}\left[a_k\right] + y_{jt} + \varepsilon_v + \varepsilon_c\right)\right), 0\right]$$

$$b_j = 0 \;\; if \;\; a_j < T, \;\; else \;\; b_j = 1$$

where δ is a decay coefficient (δ = 0.1), ι is an inhibition coefficient (ι = 0.9), ε_v is a noise component ranging over [-0.1, +0.1] and *varying* in each cycle, ε_c is a noise component ranging over [-0.2, +0.2] and *constant* for a random period ranging over [0, 10] s. Note: ε_c is very important for the actor-critic's exploration since positive and negative values of ε_v tend to sum to 0 during the races.

Posture controller This component has a 2D layer of input-units with activation z_k ($z \in Z$, $|Z|$ = 20×20 = 400). While the component is *trained* (see below), the input units are activated in [0, 1] on the basis of the (x, y) position of the arm's "hand" on the plane and on the basis of a Gaussian activation field (see Figure 1):

$$z_k = \exp\left(-\frac{\left(x_k - x\right)^2 + \left(y_k - y\right)^2}{2\sigma^2}\right)$$

where σ is the standard deviation (set to 3.33 cm) and (x_k, y_k) is the "preferential" hand position of unit k. The activation of the units is normalized to have a total activation of them equal to 1. When the controller is used to reach a target position, only one of these units is activated with 1 by the topologically corresponding winning unit of the action selector (for simplicity, only 9 units/targets are involved by this process). In a second version, the posture controller has two further input units with activation α and β ($\alpha,\beta \in [0, 1]$) that encode the current normalized angles of the arm. The two versions of the controller correspond to a different modeling of the function played by *fiber-muscle afferents* (sensors located in the muscles, such as the *Golgi tendon-organs*, that return information such as muscles' length and tendons' stiffness to the spinal cord and brain, cf. Shadmehr & Wise, 2005). All the input units (in both versions of the controller) are connected all-to-all to 2 sigmoid output units. These output units encode either the *desired angles* of the arm (first version), in which case they are denoted as α' and β' ($\alpha',\beta' \in [0.25, 0.75]$), or the *desired change of these angles* (second version), in which case they are denoted as $\Delta\alpha$ and $\Delta\beta$ ($\Delta\alpha$, $\Delta\beta \in [0, 1]$).

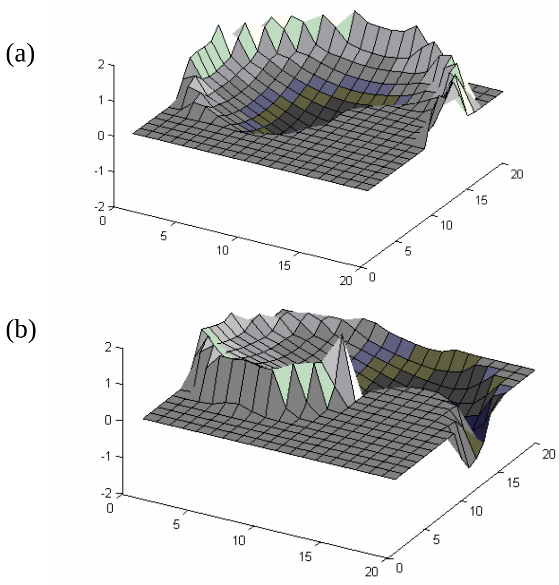

Figure 3: Weights, after training, of the second version the posture controller connecting the input map to the output units controlling the elbow (a) and shoulder (b) joints.

Childhood phase (posture controller's learning) During this phase the posture controller is trained to perform movements that allow it to reach particular target points on the horizontal plane with its "hand". The target points to reach are encoded as (x^*, y^*) Cartesian coordinates and are used to activate the input 2D map of the controller (these points are assumed to correspond to the hand's position perceived by the retina). In particular, each unit of the map activates within [0, 1] on the basis of its Gaussian receptive field. Training, mimicking self-generated experience, is based on *direct inverse-modeling* (cf. Kuperstain, 1988): the idea behind this procedure is that the system produces random movements of the arm and learns to associate the *performed action* (network's output pattern) with the *resulting position of the hand* (network's input pattern): in the future this association will allow the system to perform a suitable action when assigned a goal in terms of hand's position to reach. The detailed steps of the procedure are as follows: (1) only second version of the controller: the current angles of the arm are used to activate the two (α, β) input units; (2) both versions: a random action ($\Delta\alpha$, $\Delta\beta$) is drawn in terms of variations of the current arm's angles within [-10°, +10°] and without violating the limits of the arm's degrees of freedom; (3) both versions: the arm performs the movement corresponding to ($\Delta\alpha$, $\Delta\beta$); (4) both versions: the new angles (α^*, β^*) of the arm, and position (x^*, y^*) of the hand, are recorded; (5) first version: an error backpropagation algorithm (Rumelhart et al., 1986; learning rate 0.1) is used to train the posture controller network to associate (x^*, y^*), taken as input, with (α^*, β^*) used as desired output (in this case, during action execution the posture controller reaches the posture (α', β') that it associates with (x^*, y^*), starting from the current (α, β), through a "servomechanism" that issues commands ($\Delta\alpha$, $\Delta\beta$) the arm having maximum size of 10°); second variant: an error backpropagation algorithm (learning rate 0.1) is used to train the posture controller network to associate (α, β) and (x^*, y^*), taken as input, with ($\Delta\alpha$, $\Delta\beta$) used as desired output. These training procedures should lead the posture controller to perform movements that allow it to reach the *desired goal* (x^*, y^*), encoded in the input-unit 2D map, *from any initial posture* (α, β).

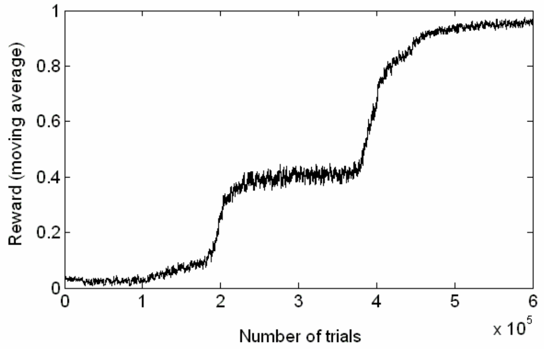

Figure 4: Moving average (1000-step window) of rewards obtained by the system during 600,000 trials of learning.

Adulthood phase (actor-critic's learning) During this phase the actor-critic component is trained (cf. Sutton and Barto, 1998) to select the correct target in the discrimination reaching task. During training R_{t+1} is set equal to 1 if the arm reaches the correct target *within 16 s. after the go signal* and 0 otherwise. Each time an action is selected and implemented, V_{t+1} is set to 0 and the trial is terminated. The *evaluator* portion of the network (i.e. the unit that produces V_t) is trained, *at each time step t* and through an error backpropagation algorithm (learning rate 0.1), to associate the following desired output to the perceived state $\boldsymbol{x}_t$:

$$R_{t+1} + \gamma V_{t+1}$$

The *actor*'s output unit of the selected action is trained, through a backpropagation algorithm and *only* in correspondence to the state $\boldsymbol{x}_t$ where the action was selected, with the following desired output (learning rate 0.4):

$$y_{jt} + (1 - y_{jt})\left|S_{t+ET}\right| \quad if\ S_{t+ET} > 0$$
$$y_{jt} + (0 - y_{jt})\left|S_{t+ET}\right| \quad if\ S_{t+ET} < 0$$

where *ET* is the duration of the execution of the action. This learning process has the following desired effects: (a) the evaluator's evaluations V_t of the perceived states $\boldsymbol{x}_t$ tend to approach γ^T, where *T* is the average number of cycles that separates $\boldsymbol{x}_t$ from the perception of *R*=1; (b) the signal sent to the action selector's accumulator unit corresponding to the executed action is increased if $S_{t+1} > 0$, in correspondence of $\boldsymbol{x}_t$, so that this action will have higher chances to win the race when $\boldsymbol{x}_t$ is encountered again, while it is lowered if $S_{t+1} < 0$; (c) the signal sent to the other actions' accumulators is not changed.

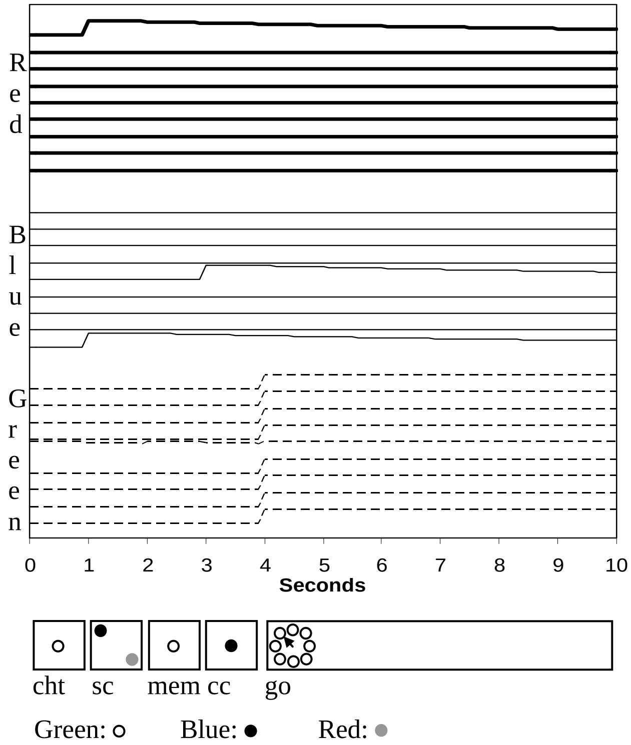

Figure 5: Top: activation, during one trial, of the 27 memory units corresponding to the 9 target positions and 3 colors (green: dashed lines; blue: continuous lines; red: bold lines). Bottom: activation of the green, blue and red cues on the screen during the trial: the boxes cover the duration time of the phases of the task (see introduction).

Results

The training of the posture controller during the "childhood learning phase" (1,000,000 cycles; learning rate 0.1) was successful: the error of the first version of the controller (output equal to desired posture (α', β')) decreased from 6.11 cm to 1.08 cm (average for 10,000 cycles) while the error of the second version of it (output equal to desired movement ($\Delta\alpha$, $\Delta\beta$)) decreased from 18.64 cm to 0.74 cm (Figure 2). For its higher performance, the first version of the posture controller has been used in the experiments with the whole system illustrated below.

Interestingly, in both versions the controller exhibits a relevant *generalization capability*: it is capable of reaching a target from any starting posture and to any reachable point. Notably, in the case of the controller's second version this requires more than one step, a condition for which it has never been trained (see Figure 2). This performance is based on the weights corresponding to the proprioception input units (α, β) and to the map input units (encoding (x^*, y^*)) that tend to have opposite signs. These weights cause the output pattern ($\Delta\alpha$, $\Delta\beta$) to becomes null only when the current posture (α, β) is such that the hand is on the desired position (x^*, y^*). In the case of the controller's first version, the weights emerged tend to encode the desired angles in an almost direct fashion (Figure 3).

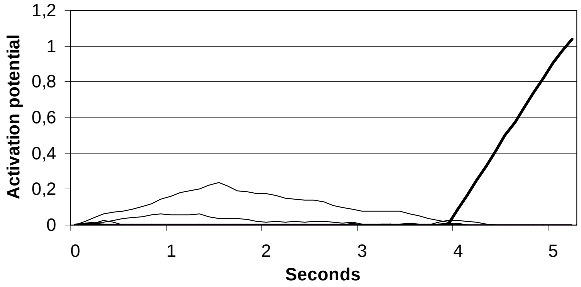

Figure 6: Accumulator units' activation potential during a race of one trial: the bold curve corresponds to the winning unit that decides the target-posture of the arm.

The training of the system (actor-critic) in the "adulthood learning phase" lasted 600,000 trials. The system learned to reach all the 16 possible targets (given by 8 positions × 2 colors) with a success of 95.78% on the last 10000 trials (Figure 4). Residual errors were due ε_c noise. Note that the two real monkeys trained by Cisek and Kalaska (2005) had a performance of 75% and 96%.

The analysis of the system's functioning shows that the memory units maintain a sustained activation during the task (Figure 5): this fuels the race of the accumulator units until one of them reaches the threshold, wins the competition, and triggers the posture controller to pursue the target posture corresponding to it (Figure 6).

During training the system progressively learns to inhibit the selection of actions in the phases before the go signal (this happens after 4 s. from the start of the trial), and to act as fast as possible, that is 1.5 s. on average after the go-signal itself (cf. Figure 7). Note that, with accumulators activated with 0 at the go-signal, the selection's maximum theoretical speed is about 1.1 s., being $\tau = 0.1$, $\delta = 0.1$, and maximum values of $y_j = 1.0$).

Conclusions and Future Work

This paper presented a model that is novel under several aspects: (a) the model proposes a first integration of an actor-critic architecture, one of the best biologically-plausible models of conditional and instrumental learning and basal ganglia, with an “accumulator model”, one of the best biologically plausible models of action selection; (b) the model integrates the action-selection component with a goal-based repertoire of actions learned on the basis of self-generated experience; (c) the model presents a working hypothesis, in the form of an integrated motor primitive-based architecture, of the possible macro-organization of the sensorimotor system of mammals. Overall the results presented, although preliminary in many respects (see below), indicate that the architecture represents a integrated working hypothesis on the overall organization of vertebrates’ motor behavior that is computationally-sound.

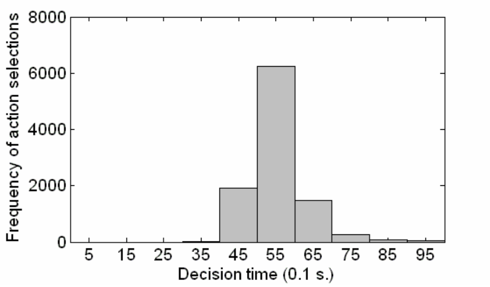

Figure 7: Histogram of reaction times of the system.

The model has also important limitations, that will be the starting point for future work: (a) the input portion of the system is a simple 27-cell map: the system should be tested with a more realistic input component; (b) the actor-critic portion of the system learns on the basis of an error backpropagation algorithm: can this be replaced by a more biologically plausible algorithm? (c) the experiments have shown that the model functions with 9 sensorimotor primitives: would it scale to larger numbers? (d) the accumulator model allows selecting only discrete and locally represented actions: is it possible to allow it to select actions represented continuously and in a distributed fashion (cf. Doya, 2000)? (e) the model does not fully takes into account empirical evidence on dopamine pathways, for example evidence on the different time courses of learning taking place in basal ganglia and prefrontal cortex (respectively fast and slow: see Pasupathy & Miller, 2005).

Acknowledgments

This research has been supported by the project “MindRACES - From Reactive to Anticipatory Cognitive Embodied Systems”, funded by the European Commission under grant FP6-511931.

References

Arbib, M. (1981). Visuomotor coordination: from neural nets to schema theory, *Cognition and Brain Theory, 4*, 23-39.

Baldassarre, G. (2002). A modular neural-network model of the basal ganglia’s role in learning and selecting motor behaviours. *Journal of Cognitive Systems Research, 3,* 5-13.

Barto, A.G., Sutton, R.S., & Anderson, C.W. (1983). Neuronlike adaptive elements that that can learn difficult control problems. *IEEE Transactions on Systems Man and Cybernetics, 13,* 835-846.

Cisek, P., & Kalaska, J. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. *Neuron*, *45*, 801-814.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation, 12*, 219-245.

Giszter, S.F., Mussa-Ivaldi, F.A., & Bizzi, E. (1993). Convergent force fields organised in the frog’s spinal cord. *Journal of neuroscience, 13(2),* 467-491.

Graziano, M.S., Taylor, C.S., & Moore, T. (2002). Complex movements evoked by microstimulation of precentral cortex. *Neuron, 34,* 841-851.

Houk, J.C.; Davis, J.L., & Beiser, D.G., (Eds.) (1995). *Models of Information Processing in the Basal Ganglia.* Cambridge, MA: MIT Press.

Joel D.E.E., Niv Y., & Ruppin E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks, 15*, 535-547.

Kandel, E.R., Schwartz, J.H., & Jessell, T.M. (2000). *Principles of Neural Science.* New York: McGraw-Hill.

Kuperstein, M. (1988). A neural model of adaptive hand-eye coordination for single postures. *Science, 239*, 1308-1311.

Nakahara, H., Doya, K., & Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences. *Journal of Cognitive Neuroscience, 13,* 626-647.

Pasupathy, A., & Miller E.K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature, 433,* 873-876.

Rainer, W.P., & Tani, J. (2004). Motor primitive and sequence self-organisation in a hierarchical recurrent neural network. *Neural Networks, 17*, 1291-1309.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533-536.

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences, 3,* 233-242.

Schall, J.D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience, 2,* 33-42.

Schultz, W., Dayan, P., & Montague, R.P. (1997). A neural substrate of prediction and reward. *Science, 275*, 1593-1599.

Shadmehr, R., & Wise, S. (2005). The computational neurobiology of reaching and pointing. Cambridge: MIT Press.

Sutton, R.S., & Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Swanson, L.W. (2005). *Brain Architecture.* Oxford: Oxford University Press.

Usher, M., & McClelland, J.L. (2001). On the time course of perceptual choice: the leaky competing accumulator model. *Psychological Review, 108,* 550-592.

von Hofsten, C. (1982). Eye-hand coordination in newborns. *Developmental Psychology, 18,* 450-461.