A REALIZATION OF GOAL-DIRECTED BEHAVIOR

Implementing a Robot Model Based on Cortico-Striato-Thalamic Circuits

Berat Denizdurduran and Neslihan Serap Sengor
Electronics and Communication Department, Istanbul Technical University, Maslak, Istanbul, Turkey

Keywords: Action Selection, Goal-directed Behaviour, Reinforcement Learning, Robot Model.

Abstract: Computational models of cognitive processes based on neural substrates clarify our understanding of the ongoing mechanisms during these high order processes. These models also inspire new approaches and techniques for implementing intelligent systems. Here, an implementation of goal-directed behaviour on Khepera II mobile robot will be presented. The main point of this work is to show the potential use of robot models for tasks requiring high order processes like goal-directed behaviour.

1 INTRODUCTION

The computational models of neural systems can be considered as tools to understand the cognition. Thus obtaining these models and showing their effectiveness would stimulate studies in cognitive science and inspire the development of new approaches for intelligent systems.

Amongst the wide spectrum of high order cognitive processes such as planning, selective attention, decision making; goal-directed behaviour has driven a specific attention. To consider goal-directed behaviour as composed of two processes: action selection and reinforcement learning bore a computationally tractable model (Sengor, 2008). Though there are numerous computational models for action selection (Gurney, 2001), (Taylor, 2000) and reinforcement learning (Schultz, 1997), (Dayan, 2009), few consider them together (Sengor, 2008), (Humphrys, 1997). These models consider the role of neural structures especially the basal ganglia, so they are biologically plausible models. Based on the models of basal ganglia behavioural disorders such as addiction (Gutkin, 2006), and different processes as feature detection are studied (Saeb, 2009).

There are also some work considering the robot models of neural substrates and some others where cognitive processes are investigated considering these robot models (Webb, 2000), (Fleischer, 2009), (Prescott, 2006). In the well-known work of Prescott (Prescott, 2006) a robot model for action selection is given. This robot model mimics the behaviour of a rat in an unfamiliar environment and it is based on mathematical model of basal ganglia which is inspired by neurophysiologic studies (Gurney, 2001). In this robot model which is implemented on Khepera II, it is shown that basal ganglia take part in selecting an action amongst different choices based on the saliencies of each possibility.

Here, the idea is to develop the work in (Prescott, 2006), further by implementing reinforcement learning to determine the saliencies which influence the choice of the rat. The process of learning has not been considered in (Prescott, 2006), where the choices depend only on a priori saliencies. So the saliencies are reconsidered and priority of one over the other is determined according to the environmental conditions with reinforcement learning. It is shown that a simpler model of the cortico-striato-thalamic circuit considered for action selection can fulfil the expected behaviour based on these saliencies. Thus, the improvement of this work over (Prescott, 2006), is the utilization of reinforcement learning to determine the choices and this is provided by using a simpler model of cortico-striato-thalamic circuit for action selection (Sengor, 2008).

In the sequel, first the computational model proposed in (Sengor, 2008) will be summarized, than the task and implementation of the model on the Khepera II mobile robot will be given. In section 3, simulation results will be given, and in the last section the expected improvements will be discussed.

Denizdurduran B. and Serap Sengor N. (2012).
A REALIZATION OF GOAL-DIRECTED BEHAVIOR - Implementing a Robot Model Based on Cortico-Striato-Thalamic Circuits.
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence*, pages 289-294
DOI: 10.5220/0003733602890294

2 COMPUTATIONAL MODEL

In this section, first a model for goal-directed behaviour (Sengor, 2008) will be introduced. Then it will be shown that the model is capable of selecting an appropriate action under changing environmental conditions and the implementation of this model on Khepera II will be discussed.

2.1 Modelling Goal-Directed Behaviour

The sub-regions of neural system communicate with each other by interconnection neurons and realize any process via neurotransmitters along neural pathways. One of these pathways is striatonigral pathway which is associated with motor control and related to dopaminergic pathway (Haber, 2010). Dysfunction of this pathway causes disorders such as Parkinson's disease, Huntington's disease and Schizophrenia (Alexander, 1990). Transmission of dopamine relates striatum with substantia nigra pars compacta. These regions are the part of the basal ganglia-thalamus-cortex circuits (Alexander, 1990). Retrograde and anterograde tracing studies have shown that the basal ganglia-thalamus-cortex circuits and a.k.a. striatonigrostrital pathways have important role in action selection and learning phenomena. The cortico-striato-thalamic model considered in this work for implementation of goal-directed behaviour is based these neurophysiological facts and is capable to explain how primates make appropriate choices and learn associations between environmental stimuli and proper actions (Sengor 2008). In (Alexander, 1990), different regions of basal ganglia are considered for different neural circuits, but principle substructures are proposed to be striatum, subthalamic nucleus, globus pallidus internal and external, substantia nigra pars reticulate and compacta. Relationship between these substructures, cortex and thalamus is very complex. The model used in this work, consider only a subgroup of these relations which are important for action selection, so it is simpler. The connections considered in the model are illustrated in Figure 1. This computational model of action selection has been shown to realize a sequence learning task (Sengor, 2008). The parameters of the dynamical system corresponding to neurotransmitters are modified with reinforcement learning. In order to realize the task, input substructure of the model which is cortex transmits the sensory data to the striatum, thalamus and subthalamic nucleus. The main effect on cortex is due to excitatory signal from thalamus.

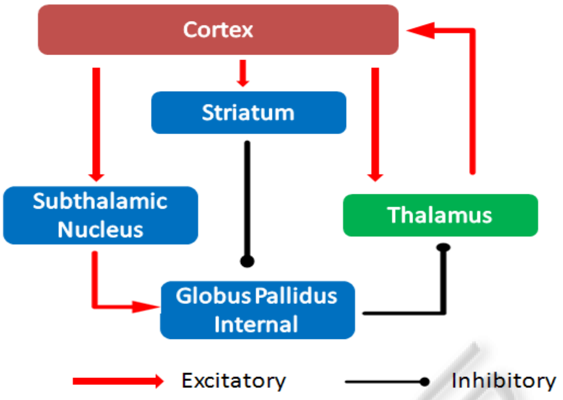

Figure 1: Basal Ganglia-Thalamus-Cortex circuit considered in the computational model.

In the model of cortico-striato-thalamic circuit for action selection (Sengor, 2008) all substructures act according to tangent hyperbolic function *f(.)*. The activity in the cortex is demonstrated by a difference equation as follows:

$$C(k+1) = f(\lambda C(k) + Thl(k) + W_c I(k)) \tag{1}$$

The variables C, Thl denote vectors corresponding to cortex and thalamus and the matrix W_c denotes the efficiency of sensory stimulus I and it is adapted through reinforcement learning. S in Figure 1 corresponds to $S = W_c I$. An action is selected, when the value of cortex variable C becomes almost one. This corresponds to firing of related neural structure.

The interconnections between substructures striatum, subthalamic nucleus and substantia nigra pars reticulate/globus pallidus interna, that are respectively denoted by Str, Stn, GP_i are modelled as in Eq. (2).

$$\begin{aligned}
Thl\ (k+1) &= f\,(C\,(k) - GP_i\,(k)) \\
Str\ (k+1) &= W_r\, f\,(C\,(k)) \\
Stn\ (k+1) &= f\,(C\,(k)) \\
GP_i\,(k+1) &= f\,(Stn\ (k) - Str\ (k))
\end{aligned} \tag{2}$$

Here W_r denotes the effect of dopamine on action selection. The action selection depends on two parameters: W_c, W_r. In (Sengor, 2008), both of these parameters were adapted through reinforcement to determine the proper action.

In this work only the effect of sensory input on action selection will be considered and W_c will be adapted. Adaptation of parameter W_c due to reinforcement learning is given in Eq. 3:

$$W_c(k+1) = W_c(k) + \mu\delta(k)S(k)C(k) \tag{3}$$

Here μ is learning rate and $\delta(k)$ corresponds to error in expectation and determined as in conventional reinforcement learning literature as follows:

$$\delta(k) = r_i + \gamma v(k+1) - v(k) \tag{4}$$

Expectation error δ depends on the value v attained to the action selected and the reward r_i and γ is discount factor. The value of the action is also updated as follows:

$$\begin{aligned} v(k) &= W_v C(k) \\ W_v(k+1) &= W_v(k) + \mu\delta(k)C(k) \end{aligned} \tag{5}$$

In the equations through 1to 5, only selecting one action is considered, so all variables are scalars. When an action has to be selected amongst a number of possible actions, except reward r_i and expectation error δ all the variables corresponding to neural substrates will be denoted by vectors and parameters by matrices.

2.2 Implementation of the Model

The computational model summarized in section 2.1 will be modified and implemented on mobile robot Khepera II. While implementing the model proposed in (Sengor, 2008) on Khepera II, first the saliencies are defined based on the sensory information obtained from mobile robot. Another modification is to consider δ as vector. The task of the robot is to mimic the behaviour of a rat's search for food in an unfamiliar environment. Here, using the mobile robot, rats' behaviour such as searching for the food, recognizing the food, and avoiding an obstacle and finding the nest is simulated. All these actions will be realized in the context of goal-directed behaviour, thus action selection based on reinforcement learning will be used. Thus behaviour of the rat will be realized mimicking the cognitive processes ongoing in neural structures.

Figure 2 gives a schema of the whole implementation. First the perception of environment through sensors is realized and then these sensory data interacts with action selection and reinforcement learning blocks to fulfil the goal-directed behaviour.

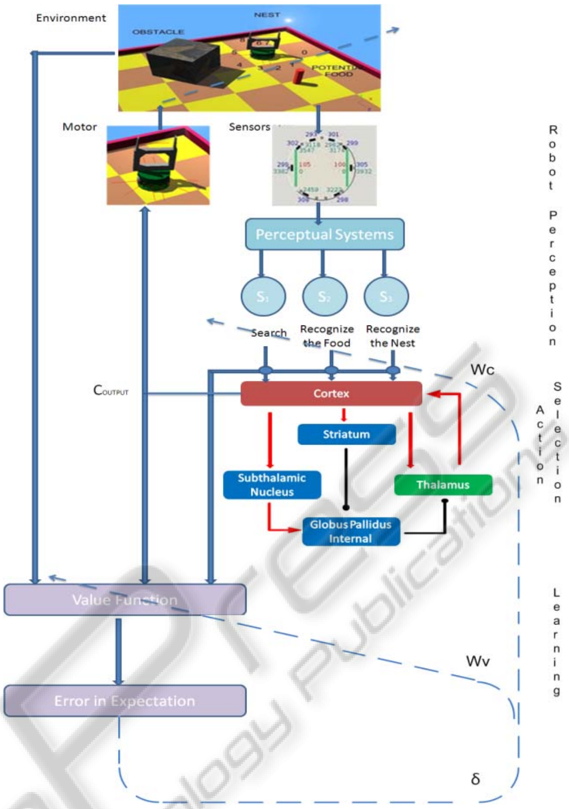

Figure 2: The architecture of the model realizing goal-directed behaviour.

Based on the structure of Khepera II mobile robot, the distance and light sensors are used to collect data from environment. As it can be followed from Figure 2, Khepera II mobile robot has 8 distance and light sensors, respectively.

These data collected from sensors form the model input vector I which is weighted by coefficient matrix W_c to define the saliencies $S = W_C I$. The dimension of this matrix is determined by the number of action choices and the saliencies build up the perceptual system. This matrix is modified through reinforcement learning process. In the problem considered, there are three saliencies corresponding to search, recognizing the food and recognizing the nest and they are formed with the data collected from sensors. This data is considered with three different aspects corresponding cylinder and nest distance and gripper position. So, the saliencies are defined as follows:

$$S = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} p_{cyl} \\ p_{grip} \\ p_{nest} \end{bmatrix} \tag{6}$$

Since there are three saliencies S built by weighting three sensor information I, the dimension of vectors and the matrix are $S, I \in R^3$, $W_C \in R^{3\times 3}$, respectively. Robot distance sensors give natural numbers between "0" to "1050" which means absence or presence of related object. These numbers are scaled in "0" to "1" so the variables corresponding to distance of cylinder, gripper position and distance of nest $p_{cyl}, p_{grip,} p_{nest}$ are denoted by rational numbers. This scaling is given in Eq. (7):

$$dist\_sens\_val(i) \doteq 1 - 0.001\, dist\_sens(i) \qquad (7)$$

Once the saliencies are established, the cortico-striato-thalamic circuit determines an action. This selected action and the reward obtained for it interacts with the learning block and the selection process in action selection block restarts at each time step. During this process, W_C is adapted continuously, till the robot comes across cylinders or nest. During the learning process, saliencies determine the action selected, and selected action is used to control the behaviour of robot. Unlike the work of Prescott et al., (Prescott, 2006) there is no need for a busy signal as the sensor data is considered constantly.

The action selection adapted by reinforcement learning block is the contribution of this work. In (Prescott, 2006), the idea of behavioural selection is based only on certain targets and the sensor data and it is designed on a rule based algorithm. Early studies of action selection and reinforcement learning phenomena are proposed in different contexts for separate tasks. However, both tasks are considered together in this work.

3 SIMULATION RESULTS

Khepera II mobile robot is used to simulate the task of a rat searching for food in an unfamiliar environment, recognizing the nest and carrying food there. We simulated rats' intrinsic feelings in a simple learning task. The robot is placed in any starting point from which the rat could do any one of the three goal actions. Only one of the goal actions is chosen on each trial, and the chosen action is searching in the beginning of the experiment. The experiment is illustrated in Figure 3 (b).

Experiments such as those illustrated in Figure 3 would clarify the difference between each process. The case in Figure 3 (a) corresponds to the work in (Prescott, 2006) where the saliencies are determined a priori. So robot begins to search instantly with the correct choice, it recognizes the obstacle and food without mistake. When food is picked up by the robot, the light sensor begins to search light source which is the indicator of the nest. Notice that for this process the search continues until light sensors recognize the nest.

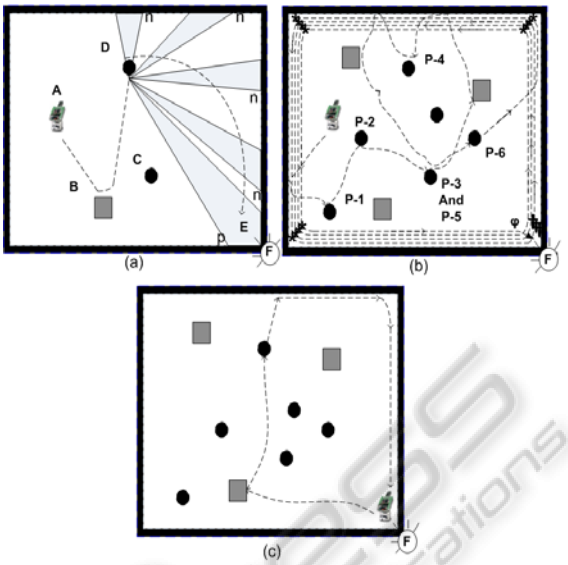

Figure 3: Robot foraging on an unfamiliar environment is illustrated with a priori saliencies, during learning and after learning, respectively in (a), (b) and (c). n: negative for light, p: positive for light, A: Khepera II, B: Obstacle, C, D: Potential food, E: Nest, F: Light, P: Process of Learning, ∗: negative for nest, †: nest but not enough for deposit, φ: deposit it to the nest.

In Figure 3(b), there are no a priori determined saliencies, the robot learns the environment with the choices it makes and the rewards it obtains. Thus it begins with random search and it takes some trial and error steps till it finds food, picks it up and carries to the nest. Once the learning process is completed, it can immediately pick up the food and carry it to nest as shown in Figure 3(c).

As the robot is not moving at the beginning of the experiment depicted in Figure 3(b), the reinforcement learning block force it to move and begin to search. This is provided by increasing coefficient "a_{11}" through reinforcement learning.

In Figure 4, the adaptation of coefficient "a_{11}", change in expectation error and reward are given, respectively. Once "a_{11}" is large enough and "search" salience is selected, robot begins to move. If Khepera II robot comes across to any one of the potential food, coefficient of "a_{22}" begins to increase. This is the learning phase of recognizing food. Results of this phase are given in Figure 5.

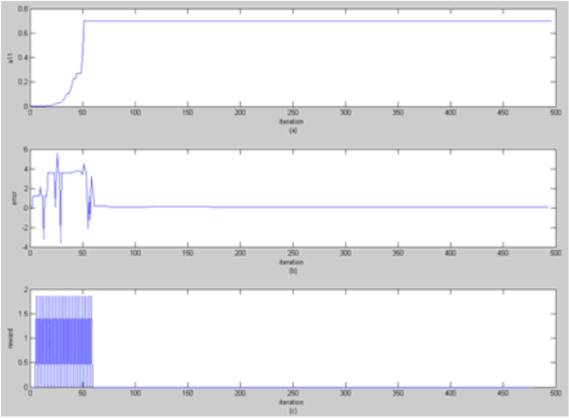

Figure 4: Simulation results for searching phase during learning process. After 80th iteration, learning process ends for the search salience but the given illustration is continued till the 500th iteration. Reward is 1.8 for this coefficient.

Once the robot learns to recognize food and picks it up, it has to begin searching the nest. In order to reach the nest the robot moves along the wall and seeks for the light source when it finds it, reward is given. This reward modifies the value of "a_{33}" and when "a_{33}" reaches a certain value, the robot learns the place of the nest. There are six coefficients besides "a_{11}", "a_{22}", "a_{33}" but these are not necessary for the determination of saliencies, so they are kept constant.

Once, the robot places the food to the nest, the search for food begins again, but as it learned the food and the nest it picks up the first food it comes across and carries it to the nest directly. As, the robot do not learn the coordinates of the latest food it picked up, the searching process is made randomly.

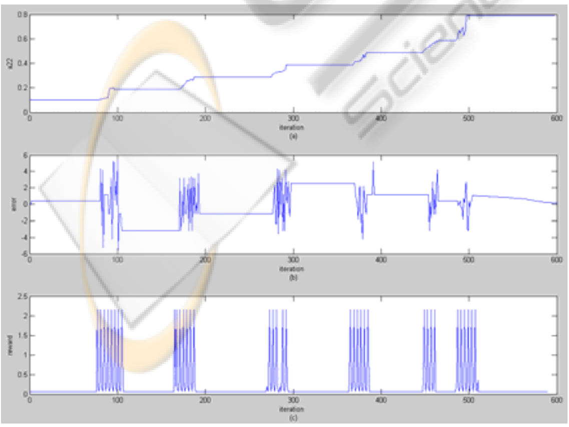

Figure 5: Simulation results for pick up and carrying phase. After 550th iteration, learning process ends for this phase. Reward is 2.2 for this coefficient.

Learning the deposit of the food in the nest is same as the other learning routines and the results are given in Figure 6. We can change the reward value for each step to examine the variation of process, so each reward selected differently. As it can be followed from Figure 4, the small reward need more iteration to learn the task, and when the reward value is increased less iteration step is needed and less try outs to learn the task. In Figure 5, the reward is chosen as 2.2 so robot learns the potential food in 6 try outs while in Figure 6 reward is increased and learning process for the nest ends in 4 try outs. This reinforcement learning results are drawn in MATLAB, but the data are collected from the environment where mobile is implemented. The mobile robot is trained to learn to recognize the food and the place of the nest and it is capable of completing the task even though the conditions in the environment changes. In 20 trials, mobile robot recognizes the food and the nest for 17 cases. Once the learning is completed the robot learns the place of the nest and deposits the food there.

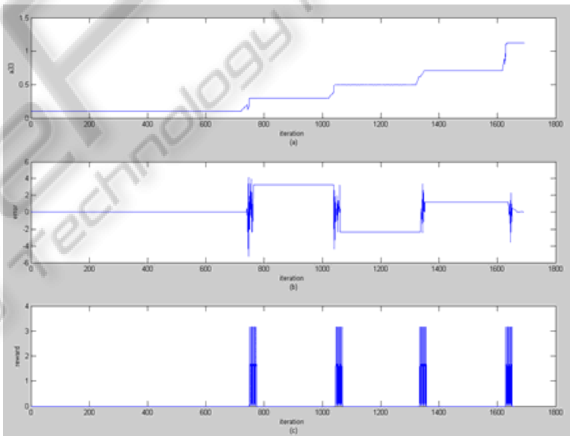

Figure 6: Simulation results for learning to find the nest and deposit phase. After 1700th iteration, learning process ends for finding the salience so in figure 1800th iteration are shown. Reward is 3 for this coefficient.

4 CONCLUSIONS

In this work, it is shown that robot implementation of neural circuits which are capable of realizing reinforcement learning is possible. Here, the model proposed in (Sengor, 2008) is reconsidered and implemented on mobile robot Khepera II to mimic the behaviour of a rat searching for food in an unfamiliar environment. It has to be emphasized that a more complex cognitive process than action selection, i.e., goal-directed behaviour is implemented on a mobile robot. So, the work considered here improves (Prescott, 2006), in two aspects, reinforcement learning process is implemented on Khepera II and goal-directed behaviour is realized. The task considered could be easily upgraded for more complex scenarios.

Here the choices of the robot are determined only by saliencies depending on sensor data. So the action selection is due to environmental inputs. In (Shultz 1997, Dayan 2009), it has been discussed that the action selection is affected also by the dopamine value which is determined by emotional processes. Thus the choices of the robot should also be determined by W_r parameter. So the adaptation of W_r could be considered to model the emotional drives.

ACKNOWLEDGEMENTS

The mobile robot Khepera II used for simulations and implementations belong to I.T.U. Artificial Intelligence and Robotics Laboratory. The authors would like to thank the laboratory staff and especially the coordinator Sanem Sarel Talay for their guidance and for their sharing knowledge. This work is partially supported by I.T.U. BAP project.

REFERENCES

Gurney, K., Prescott, T. J., Redgrave, P., 2001. Computational Model of Action Selection in the Basal Ganglia I: A New Functional Anatomy. *Biological Cybernetics*, vol.84, 401-410.

Taylor, J. G., Taylor, N. R., 2000. Analysis of Recurrent Cortico-Basal Ganglia-Thalamic Loops for Working Memory. *Biological Cybernetics,* vol.82, 415-432.

Schultz, W., Dayan, P., Montague, P. R., 1997. A Neural Substrate of Prediction and Reward. *Science* 275, 1593-1599.

Dayan, P., 2009. Dopamine, Reinforcement Learning, and Addiction. *Pharmacopsychiatry.* Vol.42, 56-65.

Gillies, A., Arbuthnott, G., 2000. Computational Models of the Basal Ganglia. *Movement Disorders.* 15, no. 5, 762-770.

Sengor, N. S., Karabacak, O., Steinmetz, U., 2008. A Computational Model of Cortico- Striato-Thalamic Circuits in Goal-Directed Behavior. *LNCS 5163, Proceedings of ICANN,* 328-337.

Gutkin, B. S., Dehaene, S., Changeux, J. P., 2006. A Neurocomputational Hypothesis for Nicotine Addiction. *PNAS*, vol.103, no.4, 1106-1111.

Saeb, S., Weber, C., Triesh, J., 2009. Goal-directed learning of features and forward models. *Neural Networks*, vol.22, 586-592.

Webb, B., 2000. What does robotics offer animal behavior? *Animal Behavior,* Vol. 60, 545-558

Fleischer, J. G., Edelman, G. M., 2009. Brain-based devices. *IEEE Robotics and Automation Magazine,* 33-41.

Prescott, T. J., Montes-Gonzalez, F. M., Gurney, K., Humpries, M. D., Redgrave, P., 2006. A Robot Model of the Basal Ganglia: Behaviour and Intrinsic Processing. *Neural Networks,* 1-31.

Haber, S. N., 2010, The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology Reviews*, 35, 4-26.

Alexander, G. E., Crutcher, M. D., DeLong, M. R., 1990. Basal ganglia-thalamocortical circuits: Parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Progress in Brain Research*, 85, 119-146.

Humphrys, M. "Action Selection Methods Using Reinforcement Learning", Ph.D. Thesis, Trinity Hall, Cambridge, 1997.

APPENDIX

The algorithm corresponding to the model considered is summarized as follows:

```
Begin
SetCoefficients
SetInitialCond
GetSensorData
ScaleSensorData
1. ReinforcementLearning
   If ∀DistSen=0&&grip=0&&wheels=0
     EvaluationOfEquation 6-10
    Update $a_{11}$
   If DistSen2&&DistSen3=0
     EvaluationOfEquation 6-10
    Update $a_{22}$
   If ∀LightSen!=0&&∀DistSen!=0
     EvaluationOfEquation 6-10
     Update $a_{33}$
2. Saliencies
   $S_i=\sum_{j=1}^{3} a_{ij}I_{ij}$;i=1,2,3
3. Action Selection
   For IterationStep<200
     EvaluationOfEquation 1-5
4. RobotMotion
   If e1>0.67&&e2<0.67&&e3<0.67
     DoSalience1
   If e1<0.67&&e2>0.67&&e3<0.67
     DoSalience2
   If e1<0.67&&e2<0.67&&e3>0.67
     DoSalience3
   // $e_i$ are cortex output values
End.
```