Study of Skinner Automaton Implemented on a Two-Wheeled Robot

Xuan Wu, XiaoGang Ruan, XiaoPing Zhang, Ouattara Sie
Institute of Artificial Intelligence and Robots, School of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
abc0767@live.com, adrxg@bjut.edu.cn, zhangxiaoping369@163.com, sie.chine@gmail.com

***Abstract*—Learning is the main aim of robotics. In this paper we present a new stochastic learning automaton called a Skinner automaton as a psychological model for formalizing the theory of operant conditioning. We identify animal operant learning with a thermodynamic process, and derive a so-called Skinner algorithm from Monte Carlo method and Metropolis algorithm and simulated annealing. The Skinner automaton is implemented on a two-wheeled robot with a flexible lumbar in a simulation experiment and it learns to keep balance successfully.**

Keywords- operant conditioning; learning automaton; two-wheeled robot

I. INTRODUCTION

One of the most important characteristics of robots is that they can adapt the unknown environments and learn by themselves. The traditional way of machine learning is directly giving the robots the knowledge. However, we cannot predict all the possibility in an unknown environment, which means we cannot give the robot the full knowledge it needs to adapt the environment. The robots need to be able to learn by themselves.

Bionic learning is being paid more attention to researchers, and operant conditioning is a learning method that can make machine learn in an animal-like way.

When an animal is in an unknown environment, it has to get information from the environment. B. F. Skinner's work has introduced Operant conditioning (OC) in 1938 [1]-[2]. Thanks to his pioneering contribution, sometimes OC is also called Skinnerian conditioning.

OC is an important form of psychological learning and sometimes called operant learning, with which humans and animals learn to associate their behaviors with the consequences. It inspires the study of machine learning.

The earlier work was contributed by Grossberg [3]-[4] whose model was used for robots to learn the operant behavior of obstacle avoidance [5].

In this paper, we represent a new psychological model called Skinner automaton (SAUTO) for formalizing the OC theory.

Based on the OC theory, we implemented SAUTO on a two-wheeled robot with a flexible lumbar so that the robot can learn balance skills by itself.

The balancing control of a robot is a common problem. As to human beings and animals, the balance skills can be gradually formed, developed and improved based on OC. In this paper, a robot can learn balance skills gradually like human beings and animals is showed. The Skinner automaton enables autonomous agents including robots to autonomously learn in an animal-like way.

II. SKINNER AUTOMATON

A SAUTO can be defined with a 6-tuple:

$$\mathrm{SAUTO} = (t,\ S,\ O,\ M,\ G,\ A), \tag{1}$$

where: $t \in \{0,1,2,\cdots\}$ represents the discrete time, S a set of internal states, O a set of operants, M: $S(t)\times O(t)\times S(t+1)\to R^{+}([0,+\infty))$ the motivated unit, G: $S(t)\to O(t)$ (p) the operant selection process, and A the Skinner algorithm.

G is the operant selection process of the SAUTO that, at each stage t, maps the current state $s(t)\in S$ into the current operant $o(t)\in O$ with the occurrence probability p of the operant $o(t)$ at $s(t)$. In other words, at each stage t, the SAUTO selects the current operant (action) based on the current state $s(t)$.

M is the motivated unit of the SAUTO, which is an energy system with the internal energy function E_S and the operant energy function E_O

$$\Delta E_s(t) = E_{SYS}(t) - E_o(t), \tag{2}$$

where $\Delta E_S(t)=E_S(t+1)-E_S(t)$ is the increment of the internal energy from t to $t+1$.

At each stage t, it maps the internal energy increment $\Delta E_S(t)$ from $s(t)$ to $s(t+1)$ and the operant energy $E_O(t)$ of the operant $o(t)$ that conduces to $\Delta E_S(t)$ into $R^{+}([0,+\infty))$. The motivated unit M desires the SAUTO to hold the lowest-energy state by the lowest-energy operant.

With the concept of reinforcement, we can imply that the probability of operants' updating scheme is as follow

$$p(o|s) = e^{-E_{SYS}(o|s)/K_B T} / Z(s), \tag{3}$$

where $\Delta p(o(t)\,|\,s(t))$ is under the state $s(t)$ the probability SAUTO choose the operant $o(t)$. K_B is Boltzmann's constant

In traditional learning automata theory, the inputs from external environment can be seen as the result of its previous behavior. While in contrast, during the OC process, we just consider the feedback from environment are only the mediate consequences instead of the real ones. Thus the result of an action is the change of the agent's internal state, or we can say that the current state of the agent is the consequence of its previous action.

In the OC theory, organisms are tending to show the actions that can get reward rather than those lead to punishments.

From psychodynamics [6]-[7] and biological thermodynamics [8], we can see that OC is a thermodynamic process, which implies that we can combine the Metropolis algorithm and simulated annealing with SAUTO.

The Metropolis algorithm is based on Monte Carlo method, which relies on repeated random sampling to compute its results [9]. It can be used to investigate the relation between operants and consequences of the SAUTO in OC. However, there is no conditioning between operants and consequences in the Monte Carlo method.

In the Metropolis algorithm [10]-[11], an internal energy function E is defined for thermodynamic systems. Their key contribution is that, instead of unconditionally accepting the random mutations s' generated from the current configuration s, they accept the s' with a probability. When $\Delta E_S \leq 0$, the new state is unconditionally accepted, or accepted by the probability $e^{-\Delta E/TK_B}$, where T is temperature, K_B is Boltzmann's constant. This means the lower the internal energy of the new state, the larger its opportunity being accepted.

In the metropolis algorithm, the new state can be rejected. In OC, however, for organisms the new state cannot be rejected because it is the consequence induced by its behavior $o(t)$. Instead, it can only try to avoid correlative operant.

Inspired by metallurgy methods, and for global optimization, Kirkpatrick, Gelatt, and Vecchi in 1983 [12], and Černý in 1985 [13] introduced simulated annealing and generalized the Metropolis algorithm by introducing a temperature scheme. Simulated annealing starts the Metropolis algorithm with a high temperature, and then, slowly cools so that the search space gradually shrinks, and lastly to a small set of states with the global energy minimum.

At the beginning, for each state, SAUTO has a large set of optional actions, and then some actions are no longer selected for they make the internal energy of the agent get higher. It means that a good action that decrease the internal energy will occur more as the temperature cools down.

We can see that both annealing and animal learning are gradual optimization processes. OC is a process of behavior optimization. The Metropolis algorithm is started by simulated annealing, temperature slowly cools while the search space shrinks gradually to a small set of actions corresponded to each state with the global energy minimum.

From above, we have the updating scheme of the occurrence probabilities of the operants

$$p(o|s) = e^{-E_{SYS}(o|s)/K_B T} / Z(s), \qquad (4)$$
$$Z(s) = \Sigma_{(o \in O)} e^{(-E_{SYS}(o|s)/K_B T)}, E_{SYS(o|s)} = \Delta E_{SYS(o|s)} + E_o(o).$$

SAUTO simulates the OC and the Skinner algorithm is derived the Metropolis algorithm and simulated annealing. The SAUTO runs like a thermodynamic system and tries to gain optimal learning with the Skinner algorithm.

The Skinner algorithm runs as Fig. 1 shows.

Initializing: set the internal energy E_{SYS}=0, set the maximum artificial temperature T_A (=$K_B T$) be large enough.

Behavior selecting: Compute the occurrence probabilities of the operants (in O) at the $s(t)$ with E_{SYS}, and select an operant o from O with the probabilities. Then implement the action to get the new state $s(t+1)$.

Updating probabilities: compute the internal state energy due to the new state, then from (4) compute the new probabilities of the operants of the previous state.

Cooling: Decrease the T_A according to a specified cooling schedule, and repeat from step "Behavior selecting" until the T_A is low enough.

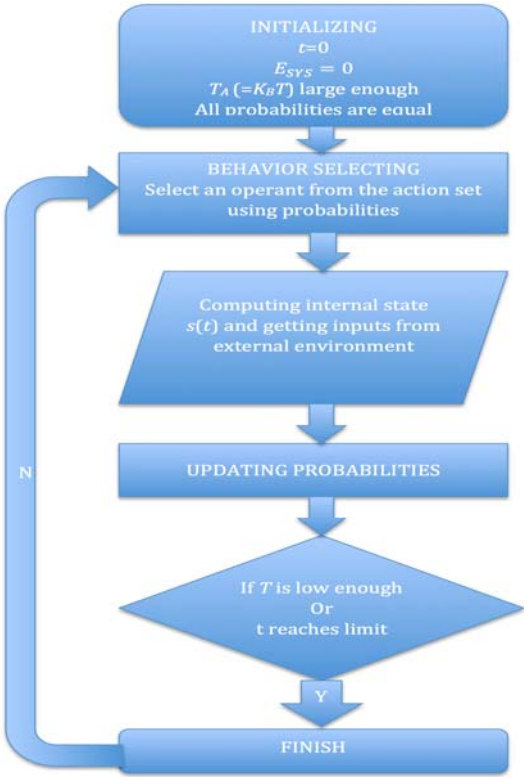

Figure 1. SAUTO's flow chart.

III. TWO-WHEELED ROBOT WITH A FLEXIBLE LUMBAR

Two-wheeled robots are a class of balancing robots, which are much absorbing due to their inherent unstable dynamics.

In a sense, two-wheeled robots are a class of bionic systems, which imitate human upright posture and attempt to exhibit some balancing skills to balance their postures. The posture balancing skills of a human are developed and shaped during operant learning, in which operant conditioning plays an important role. Without a doubt, it is significant for balancing robots to learn balancing skills like human and animals

And with a flexible lumbar will surely make balancing more difficult.

We have built a physical two-wheeled robot called Hominid 3 that is 58 cm in height and 22.5 kg in weight (see Fig. 2). Hominid 3 is a flexible balancing robot with complicated dynamics. It has a flexible lumbar made of a spring so that it is more bionic and more challenging to posture balancing. Hominid 3 has the receptor of an inertial measurement unit (IMU), and the effectors of two motors that drive the left wheel and the right wheel respectively. The IMU is used for measuring the tilt angle θ of the robot as well as the angular velocity $d\theta/dt$.

IV. SIMULATED EXPERIMENT WITH MATLAB

SAUTO serves as a sensorimotor system related posture and movement for Hominid 3 to develop its balancing skills.

The SAUTO's state is a vector of $\boldsymbol{s}=(\theta,\dot{\theta})$. The tilt angle and the angular velocity are divided into 9 levels respectively as showed in Table I., and therefore, there are 81 states in all in the SAUTO's state set. $\boldsymbol{s}$=(0, 0) is the upright state of the robot of Hominid 3. The internal energy function that represents the propensity or tropism of Hominid 3 is defined with

$$E_S=(10\theta)^2+18\theta\dot{\theta}+(2\dot{\theta})^2, \quad (5)$$

where the first item represents the propensity to regulate the tilt angle, the last with the tilt velocity, and the middle with the coupling between the tilt angle and the tilt velocity.

The SAUTO's operant that will operate on the effectors of two motors M_L and M_R is a scalar $\boldsymbol{o}=U_{PWM}$ ($\in$[−2500,+2500]), the value of Pulse Width Modulation (PWM). The PWM value is divided into 15 levels as showing in Table II, and therefore, there are 15 different operants in all in the SAUTO's operant set. The operant being selected will be operated on the servos to control the speed and direction of the wheels.

Figure 2. Physical two-wheeled robot, Hominid 3.

TABLE I. THE STATE LEVELS OF THE TILT ANGLE AND THE ANGULAR VELOCITY

Level k	States	
	$\theta\ (deg)$	$\dot{\theta}(deg/s)$
−4	(−∞, −15]	(−∞, −55]
−3	(−15, −6]	(−55, −20]
−2	(−6, −3]	(−20, −10]
−1	(−3, −0.75]	(−10, −2.5]
0	(−0.75, +0.75)	(−2.5, +2.5)
1	[+0.75, +3)	[+2.5, +10)
2	[+3, +6)	[+10, +20)
3	[+6, +15)	[+20, +55)
4	[+15, +∞)	[+55, +∞)

TABLE II. THE OPERANT LEVELS

Level k	Operants	Level k	Operants
	U_{PWM}		U_{PWM}
−7	−2500	1	+100
−6	−1850	2	+250
−5	−1250	3	+450
−4	−750	4	+750
−3	−450	5	+1250
−2	−250	6	+1850
−1	−100	7	+2500
0	0		

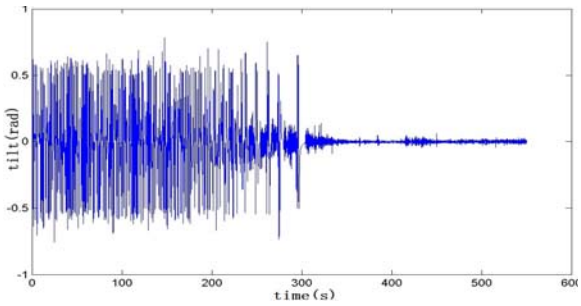

Figure 3. The tilt angle curve.

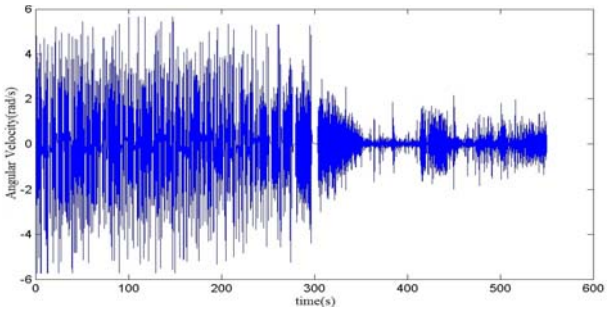

Figure 4. Angular velocity curve.

The results can be seen in Fig. 3 and Fig. 4 , SAUTO successfully learns balance skills and the robot can keep upright

From the results we can see that SAUTO learns gradually and eventually makes the robot keep upright. The learning process takes about 300s in the simulation.

Fig. 5 shows the difference between the U_{PWM} of SAUTO and PID method with a contour figure. The axis is the state of the state ***s***. The areas with different colors can clearly show the values of U_{PWM}'s distribution. And from the figures it's obvious that the SAUTO's selection scheme is almost the same with the PID method.

To make the results of the two methods more alike, we just to make more levels of the states.

V. CONCLUSION

In this paper we present a new learning automaton based on the OC theory called SAUTO, and we used it in a simulated experiment on a two-wheeled robot with a flexible lumbar. The result shows that SAUTO can gradually learn the balance skills and lastly keep the robot upright. SAUTO needs not to know the model of the agent, instead it only needs to know the levels of the states feedback from environment and the levels of actions. Next we are planning to use the SAUTO on real robot.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (No.61075110; No. 60774077; No.61375086), China's 863 Program (No. 2007AA04Z226), Beijing Natural Science Foundation (No. 4102011), and Key Project (No. KM200810005016; No.KZ201210005001) of S&T Plan of Beijing Municipal Commission of Education. National Basic Research Program of China(973 Program) (2012CB720000). Specialized Research Fund for the Doctoral Program of Higher Education (No.20101103110007).

REFERENCES

[1] B. F. Skinner, The Behavior of Organisms. New York: Appleton-Century-Crofts, 1938, pp. 61–116.

[2] B. F. Skinner, Science and Human Behavior. New York: Macmillan, 1953, pp. 45–128.

[3] S. Grossberg, "On the dynamics of operant conditioning,". J. Theor. Biol. Amsterdam, vol. 33, pp. 225–255, November 1971.

[4] S. Grossberg, "Classical and instrumental learning by neural networks," in Progress in theoretical biology, R. Rosen and F. Snell, Eds. New York: Academic Press, 1974, pp. 51–141.

[5] C. Chang and P. Gaudiano, "Application of biological learning theories to mobile robot avoidance and approach behaviors," Advs. Complex. Syst. Singapore, vol. 1, pp. 79–114, March 1998.

[6] M. Horowitz, Introduction to Psychodynamics – A New synthesis. New York: Basic Books, 1988, pp. 17–240.

[7] E. W. Brucke, Lectures on Physiology. Vienna: Braumuller, 1874.

[8] D. Haynie, Biological Thermodynamics. Cambridge: Cambridge University Press, 2001, pp. 293–330.

[9] J. von. Neumann, "Various techniques used in connection with random digits," Appl. Math. Ser. Washington D.C., vol. 12, pp. 36–38, Mar 1951.

[10] N. Metropolis, A. E. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," J. Chem. Phys. New York, vol. 21, pp. 1087–1092, Jun 1953.

[11] W. L. Jorgensen, "Perspective on 'Equation of state calculations by fast computing machines'," Theor. Chem. Acc. Berlin, vol. 103, pp. 225–227, February 2000.

[12] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," Science, Washington D.C., vol. 220, pp. 671–680, May 1983.

[13] V. A. Černý, "Thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm," J. Optim. Theory. Appl., New York. vol. 45, pp. 41–51, Janurary 1985.

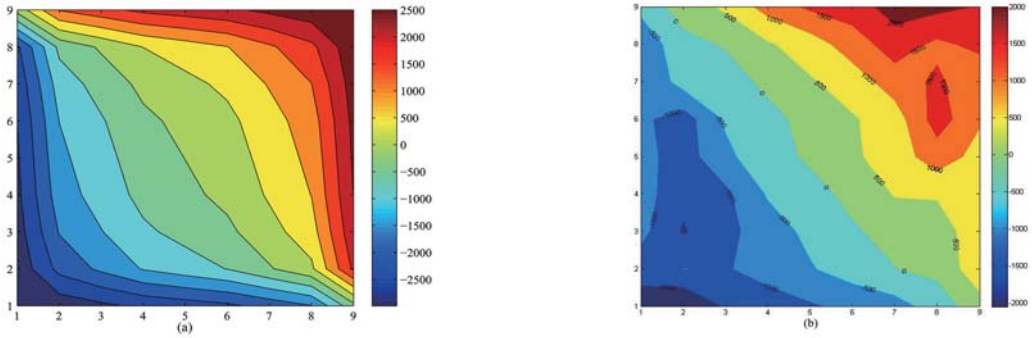

Figure 5. U_{PWM} of SAUTO and PID method with a contour figure. (a) is the contour figure of U_{PWM} of PID method, (b) is the contour figure of U_{PWM} of SAUTO. The two figures show the difference of the control scheme between PID and SAUTO.