35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Bio-robots Automatic Navigation with Graded Electric Reward Stimulation based on Reinforcement Learning

Chen Zhang[1] Chao Sun[2] Liqiang Gao[3] Nenggan Zheng[4] Weidong Chen[5] Xiaoxiang Zheng[6]

Abstract— **Bio-robots based on brain computer interface (BCI) suffer from the lack of considering the characteristic of the animals in navigation. This paper proposed a new method for bio-robots' automatic navigation combining the reward generating algorithm base on Reinforcement Learning (RL) with the learning intelligence of animals together. Given the graded electrical reward, the animal e.g. the rat, intends to seek the maximum reward while exploring an unknown environment. Since the rat has excellent spatial recognition, the rat-robot and the RL algorithm can convergent to an optimal route by co-learning. This work has significant inspiration for the practical development of bio-robots' navigation with hybrid intelligence.**

I. INTRODUCTION

Bio-robot is a biological-artificial hybrid system which the animals controlled by outer devices such as electric stimulation through Brain Computer Interface (BCI) [1]. Through mild simulation on specific regions of animals' brain, the controlling commands manipulate the animals behaviors directly [2]. BCI-based robots have been realized on different kinds of animals, for example, rats, sharks and insects [3], [4], [5], [6]. Guided by human operators, the animals can accomplish some complex tasks like walking along complicated 3-D routes. It provides a promising perspective such as searching and rescuing in disaster areas and even at battle fields. Therefore bio-robots have attracted more and more attention from researchers in neurobiology, biomedical engineering, computer science, etc.

Since the first BCI-based rat-robot navigation system was developed by S.K. Tawlar in 2002 [4], four behavioral control commands are proposed: **Forward**, turning **Left** and **Right and even Stop** [7] through delivering appropriate electrical stimulus at the rat brain. However, current researchers mainly treat bio-robots as traditional robots. They usually apply the control methods of mechanic robot directly on the bio-robots, ignoring animals' intelligence totally. These approaches perform the automatic navigation on bio-robots far less ideal, which limits practical potential heavily.

The animals learning ability to exploring the un-configured environment can be considered as a reinforcement learning process [8]. Reinforcement Learning (RL) [9] is a skill-learning process for an agent, either a mechanical device or an animal, which must change its behavior to achieve some goal optimally through interactions with dynamic environment and by awarded various forms of rewards. In recent decades, bio-inspired RL has been widely used in the mobile robot navigation for example path planning and obstacle avoidance in dynamic environment [10], [11], [12], [13]. However, most of these methods mainly concentrates on the algorithms about obstacle avoidance and prone to meet the challenge of local minima problem. Animals' excellent locomotive ability and intelligence in spatial recognition will be a perfect enhancement for reinforcement learning.

In RL navigation methods, the controlling policy determines the behavioral sequence of the robot agent. Nevertheless no policy could solve various environments in any cases subject to the complexity of the surroundings. Meanwhile some animals, e.g. rodents can adapt different cases given by proper motivation. To attain the desired reward, such as food or water, rats adopt trial-and-error strategy and find optimal route after exploration in an unknown environment. This spatial recognition ability of animals combined with the appropriate reward in RL would provide a hybrid intelligence system with the biological and artificial intelligence to implement the bio-robot automatic navigation.

In this paper, a classical reinforcement learning method, Q-learning, is introduced into the bio-robots navigation problem. The rat-robot and the rewarding algorithm working together to build up a Q-learning reinforcement procedure to meet the requirements in the unknown space. According to Q-values generated in the RL algorithm, the incremental sequence of electrical reward is given to assist the rat-robot to perceive the environment and learn the target. When the rat-robot approaches closer to the destination, it receives higher electrical stimulation corresponding to the higher Q-values, otherwise the lower stimulus. In this way, the rat-robot builds a cognitive map about the environment and learned an optimal route autonomously. It provides a practical method for rat-robot navigation in un-configured environment, effect of which is proven by the experiment.

The remainder of this paper is organized as follows. The details of our controlling algorithm in the rat-robot navigation will be described in Session II and the results will be introduced and discussed in session III. Finally, the conclusion and our future plan will be presented in session IV.

*This work was supported by grants from the 973 Program (No. 2013CB329506), the National Natural Science Foundation of China (No. 61003150, 61031002), the Fundamental Research Funds for the Central Universities.

[1]Chen Zhang, [2]Chao Sun, [3]Liqiang Gao, [4]Nenggan Zheng, [5]Weidong Chen, [6]Xiaoxiang Zheng are with Qiushi Academy for Advanced Studies (QAAS), Zhejiang University, Hangzhou 310027, China. [4]Nenggan Zheng is the corresponding author (e-mail: qaas@zju.edu.cn).

[6]Xiaoxiang Zheng is also with the Department of Biomedical Engineering, Zhejiang University, Hangzhou 310027, China. [1]Chen Zhang, [2]Chao Sun, [3]Liqiang Gao are also with College of Computer Science, Zhejiang Univerisity, Hangzhou 310027, China.

II. METHODS AND MATERIALS

A. Subjects and Reward

The navigation experiments are performed with adult Sprague Dawley rats (230-340g). One pair of bipolar electrodes is implanted in left and right medial fore-brain bundle (MFB) area (AP -3.8, ML+1.6, DV+8.2). The electrical stimulation in MFB generates intensive excitement for rats as a **Reward** command. After a 7-days recovery, the rat is placed in an operant chamber undergoing the bar-pressing training [7]. Through this traing, the rat's reward-seeking behavior and the effect of the electric stimulation is reinforced. Note that in rat-robot, only one side of MFB stimulation is chosen as the commands of Reward according to performance in the operant-chamber-tests. The difference of the side of MFB stimulation will influence the habit of rat-robot in walking. The rat-robot is shown in Fig.1.

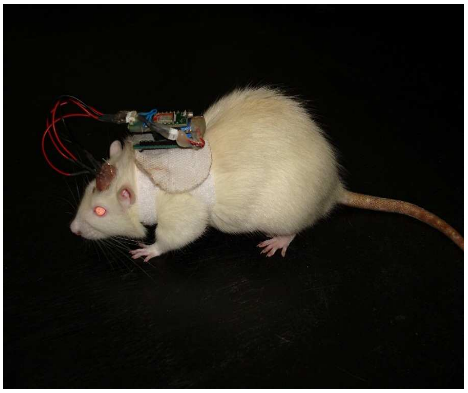


Fig. 1. The rat-robot.

B. Reward grading

Whether the rat could distinguish and react to different intensity of electrical stimulation is essential to the effect of RL methods. We grade the reward stimulation to different levels measured by rat-robots responding behaviors.

According to our experience of former navigation experiments [3], [14], [7] the parameters for MFB is given as following: pulse interval: 10ms, pulse duration: 1ms, pulse number: 10-15, pulse amplitude: 1-10V. First, the pulse amplitude starting at 1V is given to the rat and increasing in steps of 0.5V with pulse number 10, until the rat-robot can perform navigation behaviors and complete the task. Thus we get the initial stimulus parameters (labeled as n and a in TABLE I) as Level 1 which represents the lowest stimulation to guarantee the normal advance of the rat-robot. From Level 1 to Level 10 the reward variation in pulse number and amplitude is shown in TABLE I. Next the rat-robot proceeds to maze training, which consists of three trials. The distance of each trail is 420cm and the time to complete each trial is recorded. The movement speed under different stimulation levels is calculated to measure the effect of the graded reward.

TABLE I

REWARD GRADING TABLE

Level	Num	Amplitude
1	n	a
2	n	a + 0.25
3	n	a + 0.50
4	n	a + 0.75
5	n + 1	a + 1.00
6	n + 1	a + 1.25
7	n + 1	a + 1.50
8	n + 1	a + 1.75
9	n + 2	a + 2.00
10	n + 2	a + 2.25

C. Q-learning

In a standard model of Reinforcement Learning, an agent performs an action at state st with an immediate reward rt. Through this process, an action sequence is generated that enables the rat-robot to select the optimal action at any given state autonomously. In this paper, we choose Q-learning algorithm as it has been successfully utilized for solving complex RL in realistic systems with a simple rule. In Q-leaning, the Q-value $Q(s, a)$ is defined as a numerical value that evaluates the future influence on the total task when taking action a at current state s. The purpose of Q-leaning is to employ a policy π that helps the agent to select an action a that makes the Q-value maximum at a given state s. The Q-value is renewed as follows [15] :

$$Q(s, a) = r(s, a) + \gamma \max Q(\delta(s, a), a') \tag{1}$$

Where r denotes the immediate reward; δ is the state transfer function and $\delta(s, a)$ denotes the expected next state of agent after employ action a at current state s; γ is the discount rate ($0 \leq \gamma < 1$) denotes the relative ratio between the immediate reward at the current state s and the delayed reward at future state $\delta(s, a)$.

D. Application of Q-learning in rat-robot

The Q-learning algorithm can be used to provide the graded reward so that the rat-robot can learn the optimal route autonomously. There are three key points to realize this application: **States**, **Actions** and **Rewards**.

- **States:** The state in our model is represented directly by the location of the rat-robot in the maze. To simplify the controlling algorithm, these locations are divided into several sub-areas as the final states. Two special state, start and goal are assigned as the start and destination of navigation. During the experiments, a bird-eye camera and machine vision algorithms are employed to identify the instantaneous state.
- **Actions:** While in the navigation of mobile robots, the action is chosen by the controlling policy, the action is determined by the rat-robot in our experiments. The rat would choose its actions by itself according to the different reward.

- **Rewards:** We associated the graded stimulus intensity with different Q-values so that the rat-robot can receive graded rewards based on the Q value table. The distribution of the state and reward in the environment also fits the nature of spatial recognition as the *cognitive map* in the rat-robots' brain.

The whole navigation model is shown as Fig. 2. At beginning, all the items of the Q-value table is initialized to 0. Let the rat-robot walk with Level 1 reward and we record its action at every state. Once the rat reaches the goal, the Q-value of every recorded state and action is updated reversely. At the same time, the reward map is also renewed base on the Q-value table so that the rat-robot will receive graded stimulation in later trials. In the process of navigation, the Q-value table converges gradually with each complete trial. At the same time the rat learns the graded rewards under different actions and gradually tends to chose better route with higher level reward. Eventually, the Q-value table remains constant and the rat-robot always walks along the optimal route. Then we can conclude that the convergence of the RL algorithm and the rat-robots learning process of the optimal route are completed.

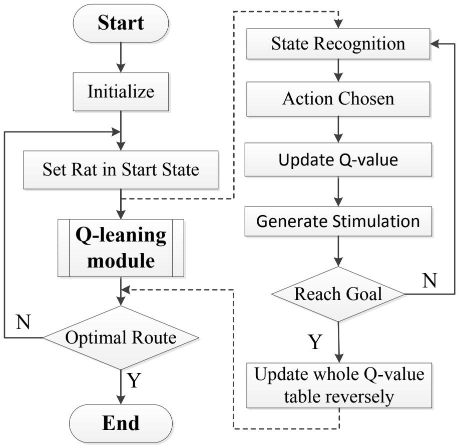


Fig. 2. Navigation model

III. EXPERIMENT AND RESULTS

A. *Reward grading in T-maze*

Rat-robot F05, V03 and V07 were tested in T-maze. The pulse parameters of the Level 1 are listed in Table II. The stimulus parameters of Level 2 to Level 10 can be calculated by Table I. Average speed of three rat-robots under different stimulation level was computed and is shown in Fig.3. The reverse-U curves illustrate the rise of the speed before Level 6. The higher intensity of the electrical reward is given, the more quickly the rat-robot response. However, after a peak at Level 6, further stimulation would cause the abnormal behaviors, such as twitching or turning around. This phenomenon also explains the descend of the speed after Level 6. Therefore in our RL navigation, the parameters from Level 1-6 are chosen in later experiments.

TABLE II
PULSE PARAMETER OF LEVEL 1

Rat No	Number	Amplitude(V)	Interval(ms)	Duration(ms)
F05	13	3.35	10	1
V03	10	4.00	10	1
V07	10	3.20	10	1

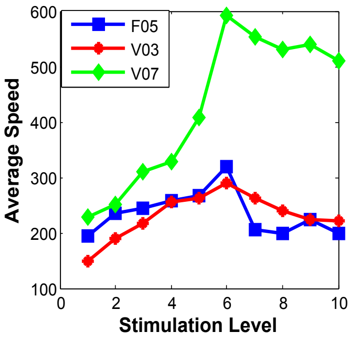
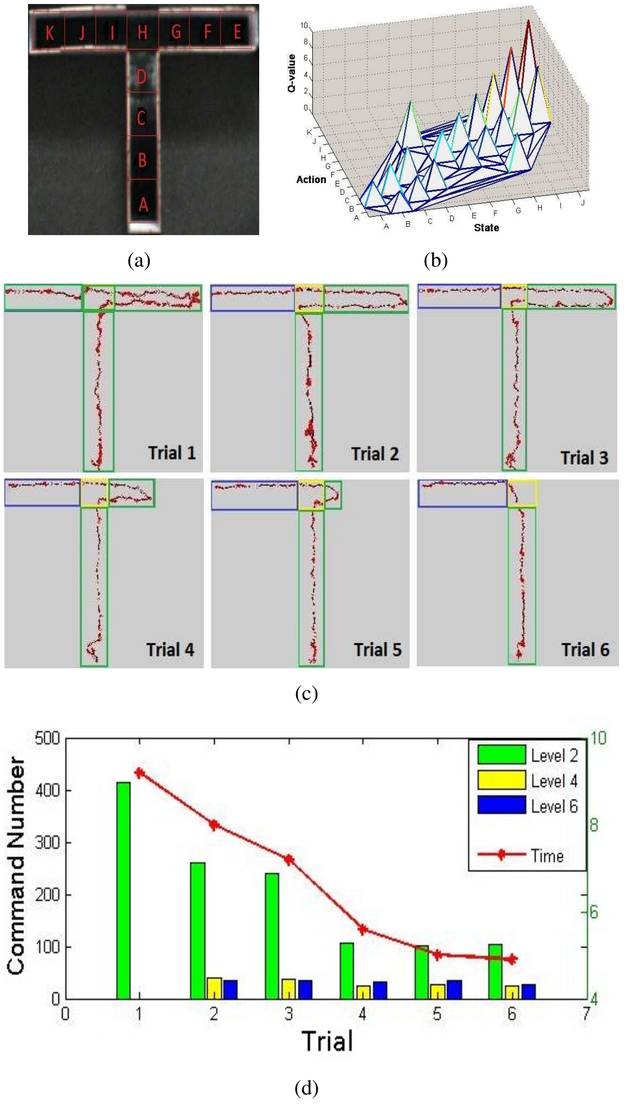


Fig. 3. Average speed in response to graded stimulus delivered to the MFB of rat-robot F05, V03 and V07.

B. *Automatic navigation with Q-learning algorithm*

The automatic navigation with graded rewards base on Q-leaning algorithm was performed in a T-maze. As shown in Fig. 4(a), we divided the T-maze into eleven states (from A to K). The rat-robot is placed at the state A and allocates state E or state K as a goal state.

As mentioned before, MFB stimulation is given in one-side of the brain, so the electrical stimulation influences the locomotive behaviors of the rat-robot [7]. Normally the rat-robot prefers to walk forward the contralateral direction of the stimulation. As in Fig. 4(a), we chose state K as goal state for Rat V07 which receives the Reward stimulation in the Left side MFB. Initially, the rat-robot preferred to turn right. Once the rat-robot reaches the goal, the Q-vale table is updated referring to the RL algorithm. In experiments, the Q-values are divided into three ranks corresponding to three kinds of rewards as Level 2, Level 4 and Level 6, which also denoted by the Green, Yellow and Blue rectangles in Fig. 4(c). All routes of six trials are shown in Fig. 4(c), with red points represent the rat-robots position at the time of receiving stimulus reward. In Trial 1, the Q-table is initialized as 0, thereby the rat-robot receives the lowest reward (Level 2) to keep walking forward freely. In later trials (Trial 2 to Trial 6), the rat-robot receives graded stimulus reward as the Q-value table updated when the rat reached the goal at the end of Trial 1. The completion time and the command number of Level 2 in Trial 1, 2, 3 descend obviously, as shown in Fig. 4(d), which certifies that the rat-robot is learning and getting familiar with the environment. In Trial 4 and Trial 5, when the rat-robot chooses the wrong direction, it perceives the lower stimulation (from Level 4 to Level 2) and then turns back to the former state for a higher reward. Eventually, the rat-robot walked along the optimal route in Trial 6. In the whole learning process, Fig. 4(b) demonstrates the convergent Q-value table. Meanwhile Fig. 4(d) shows that the completion time and commands number of each Trial descend and tend to stay stable. After 6 trials, the RL navigation algorithm and the learning process are both converged.



(a) (b)

(c)

(d)

Fig. 4. Automatic navigation in T-maze of Rat-robot V07. (a) A T-maze with eleven states: for V07, State A is the start state and State K is the goal State. (b) Results of Q-learning: Q-value of every state and feasible action is generated. (c) Actual automatic navigation route: with Q-learning process of V07 in successive six trials. Red points represent the rat position at the time of stimulation. Among this, the reward that the rat-robot receives at red points in Green Rectangle is Level 2. Similarly, Yellow Rectangle corresponds to Level 4 and Blue Rectangle corresponds to Level 6. (d) Command Numbers and Completion Time: command numbers of each level and completion time in six trials.

IV. CONCLUSION AND FUTURE WORK

A RL based control model for rat-robot automatic navigation is presented in this paper. Through maintaining a Q-value table, the best action for every state can be confirmed. Also, graded stimulation intensity is generated corresponding to the different Q-value. The results show that the rat-robot can perceive stimulation of different levels and learn the optimal route to the destination.

Our work focuses on applying RL in rat-robot automatic navigation. Based on current study, some work should be done in the future. More complicated environment should be tested. The convergence of both the RL algorithm and the learning curve of the rat depend on their learning abilities and the complexity of the environment. This process should be modeled in the future studies.

ACKNOWLEDGMENT

The authors would thank Yi Qu, Jiacheng Zhang for the rats surgery and Chaonan Yu, XinLu Zhang for helping training rats.

REFERENCES

[1] M. Nicolelis, “Brain machine interfaces to restore motor function and probe neural circuits,” *Nat. Rev. Neuroscience*, vol. 4, pp. 417–422, 2003.

[2] R. Holzer and I. Shimoyama, “Locomotion control of a bio-robotic system via electric stimulation,” in *Intelligent Robots and Systems, 1997. IROS ’97., Proceedings of the 1997 IEEE/RSJ International Conference on*, vol. 3, pp. 1514 –1519 vol.3, sep 1997.

[3] Y. Wang, X. Su, R. Huai, and M. Wang, “A telemetry navigation system for animal-robots,” *Robot*, vol. 28(2), pp. 183–186, 2006.

[4] S. K. Talwar, S. Xu, E. S. Hawley, S. A. Weiss, K. A. Moxon, and J. K. Chapin, “Behavioural neuroscience: Rat navigation guided by remote control,” *Nature*, vol. 417, pp. 37–38, May 2002.

[5] W. Gomes, D. Perez, and J. Catipovic, “Autonomous shark tag with neural reading and stimulation capability for open-ocean experiments,” *Eos Trans*, vol. 87, p. 36, 2006.

[6] X. Ye, P. Wang, and J. Liu, “A portable telemetry system for brain stimulation and neuronal activity recording in freely behaving small animals,” *Journal of Neuroscience Methods*, vol. 174(2), pp. 186–193, 2008.

[7] C. Sun, X. Zhang, N. Zheng, W. Chen, and X. Zheng, “Bio-robots automatic navigation with electrical reward stimulation,” in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 348 –351, 28 2012-sept. 1 2012.

[8] I. Whishaw and B. Kolb, *The Behavior of the Laboratory Rat: A Handbook with Tests*. Oxford University Press, 2004.

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: an Introduction*. MIT Press, 1998.

[10] L. J. Lin, “Scaling up reinforcement learning for robot control,” in *International Conference on Machine Learning*, pp. 182–189, 1993.

[11] H. R. Beom and H. S. Cho, “A sensor-based navigation for a mobile robot using fuzzy logic and reinforcement learning,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 25, pp. 464 –477, mar 1995.

[12] W. Smart and L. Pack Kaelbling, “Effective reinforcement learning for mobile robots,” in *Robotics and Automation, 2002. Proceedings. ICRA ’02. IEEE International Conference on*, vol. 4, pp. 3404 – 3410 vol.4, 2002.

[13] K. Macek, I. Petrovic, and N. Peric, “A reinforcement learning approach to obstacle avoidance of mobile robots,” in *Advanced Motion Control, 2002. 7th International Workshop on*, pp. 462 – 466, 2002.

[14] Y. Zhang, C. Sun, N. Zheng, S. Zhang, J. Lin, W. Chen, and X. Zheng, “An automatic control system for ratbot navigation,” in *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int’l Conference on Int’l Conference on Cyber, Physical and Social Computing (CPSCom)*, pp. 895 –900, dec. 2010.

[15] C. Watkins and P. Dayan, “Q-learning,” vol. 8, no. 3-4, pp. 279–292, 1992.