Proceedings of the 2006 IEEE
International Conference on Information Acquisition
August 20 - 23, 2006, Weihai, Shandong, China

Q-Learning Based Method of Adaptive Path Planning for Mobile Robot

Yibin Li
School of Electrical and Automation Engineering
Tianjin University
Tianjin, china
School of Control Science and Engineering
Shandong University
Jinan, Shandong province, china
liyb@sdu.edu.cn

Caihong Li and Zijian Zhang
School of Computer Science and Technology
Shandong University of Technology
Zibo, Shandong province, china
School of Control Science and Engineering
Shandong University
Jinan, Shandong province, china
lich@sdut.edu.cn

Abstract **- Reinforcement learning (RL) is a learning technique based on trial and error. Q-learning is a method of RL algorithms. It has been applied widely in the adaptive path planning for the autonomous mobile robot. In order to decrease the learning space and increase the learning convergent speed, this paper adopts Q-layered learning method to divide the task of searching optimal path into three basic behaviors (or subtasks), namely static obstacle-avoidance, dynamic obstacle-avoidance and goal approaching. Especially in the learning for the static obstacle-avoidance behavior, a novel priority *Q* search method (*PQA*) is used to avoid the blindly search of the random search algorithm (*RA*) which is always used to select actions in Q-learning. *PQA* uses the sum of weighted vectors pointing away from obstacles to predict the magnitude of the reinforcement reward receiving from the possible state-action after executing the action. Robot controller will select an action based on the result at the next executing time. At last *PQA* and *RA* are both simulated in two different environments. The learning results show that learn steps are fewer by *PQA* than by RA under same environment to achieve the task. And in the total learning periods *PQA* has the higher task complete percent. *PQA* is an effective way to solve the problem of the path planning under dynamic and unknown environment.**

Index Terms - Q-learning, adaptive path planning, mobile robot, PQA, RA

I. Introduction

Autonomous mobile robot can execute complex tasks in a robust manner and with minimal human supervision under uncertainty environment through sensing external surroundings by various sensors. But now the mobile robot lack flexibility and autonomy. Most of them execute predetermined actions in the high structured environments. When lies in a new circumstance or encounters an unexpected instances, it will fail to achieve the task due to the unstructured and uncertainty of the real environment. The key problem is that the robot should have the ability to learn and adapt to the environment. Adaptive path planning is one of the most important problems of the robot research in unknown and dynamic circumstance. It should satisfy some optimal criterion and find a collision-free path from the start and to the goal through the interaction between the mobile robot and the environment. To solve this problem now people adopt the behavior-based path finding method [1][2] which is influenced by the brooks behaviorism [3]. It maps the sensor information, such as the external surroundings and the internal state into the moving velocity and the direction of the robot. Furthermore the mapping relationship is achieved by learning.

There are three main learning methods, namely reinforcement learning [4], artificial neural network [5] and evolutionary algorithm [6]. Reinforcement learning allows robot to accomplish the task by learning without the need to prescribe all actions beforehand. It is a novel learning method based on the dynamic planning and supervised learning. It gets the rewards and punishment by the trial-and-error interacting with the environment to ameliorate the performance of the robot. It has been a very useful method due to its prominent ability to solve the complex problem. Its goal is to find an optimal mapping relationship between the state space and the action space. In the real environment, mobile robot has a group of discrete or consecutive states and actions. To describe all of them need a large memory even to a simple robot. This forms the combination explosion problems in reinforcement learning. To solve this problem, we adopt Q-layered learning method [7] of the reinforcement learning in this paper. Decrease the combination numbers by dividing the solution space into three subspaces. In order to increase the convergent speed of the learning method the path finding problem is divided into three subtasks (or behaviors), namely static obstacle avoiding, dynamic obstacle avoiding and goal approaching. Furthermore a novel prior *Q* search method (*PQA*) is used in the static obstacle avoiding behavior. It predicts the reinforcement signal after executing the possible state-action pair through a sum of weighted vectors pointing away from obstacles and uses it as an evidence to select next action. So the robot can visit the actions which have a high accumulative rewards frequently. In turn the learning speed can be increased. At last the method is simulated. The simulation result shows that the method can convergent quickly and have a good performance. Behavior learning

II. Behavior Learning

Behavior-based robots choose suitable behaviors and generate actions according to the local information of the environment. It is based upon the idea of providing the robot

with a collection of simple basic behaviors. The global behavior of the robot emerges through the interaction between those basic behaviors and the environment in which the robot finds itself [8]. In this paper, The behavior learning of path planning includes two tasks, one is the learning for the three basic behavior, namely static obstacle avoiding, dynamic obstacle avoiding and goal approaching. The other is the advance behaviors coordination learning. Resolve the optimal path based on the coordination of the three basic behaviors. Although the task is different for learning of each behavior, whatever behavior is, it is to learn the mapping relationship between input states and output actions, namely solving the moving velocity and direction of the mobile robot at next time according to the current environment information and interior state of the robot measured by sensors. The structures for mobile robot agent using the Q-layered method are illustrated in Fig.1 and Fig.2. Fig.3 illustrates a reinforcement learning model for the mobile robot.

Q-learning is one kind of reinforcement learning methods. It uses *Q(s,a)* to indicate the prediction values for each state-action pair. Where *s* is state and *a* is the possible action. Here, $s \in S$, $a \in A$, S is the state set and A is the action set. *Q(s,a)* is the maximum discounted reward that can be achieved by starting at state *s*, taking an action *a*, and following the optimal policy thereafter. It can be described as: $\pi^* = \arg\max_a Q(s,a)$, which π is the mapping relationship between the state and action, namely $\pi : S \to A$. *Q(s,a)* is updated by interacting with the environment. The problems of the dynamic obstacles avoidance, goal approaching and behavior coordination are not introduced in this paper. Consult the reference [4] please. In this paper we mainly research the basic behavior learning of the static obstacles avoidance.

Application of Q-learning should know the environment information, the current state of the robot and the executable action space of the motor. In order to achieve the smooth trajectory and satisfy some moving requirement of the robot, the action space are divided into six discrete motions, namely 0^0, -20^0, -40^0, $+20^0$, $+40^0$ and -180^0 angle with reference to the moving direction of the robot and moves some distance forward. External environment information S_{ext} are obstacles

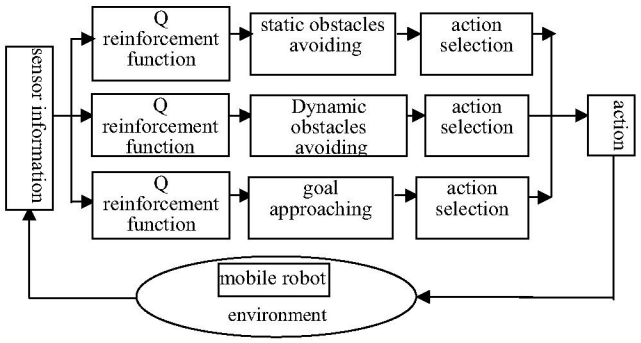

Fig. 1 Basic behaviors learning structure

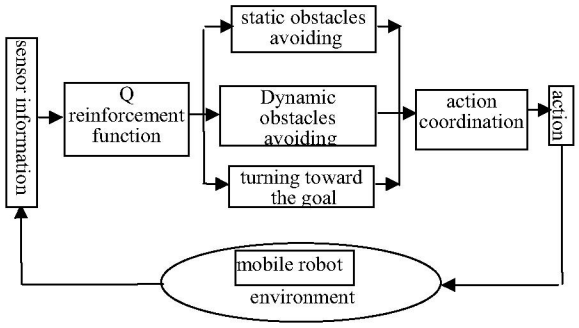

Fig. 2 Behaviors coordination learning structure

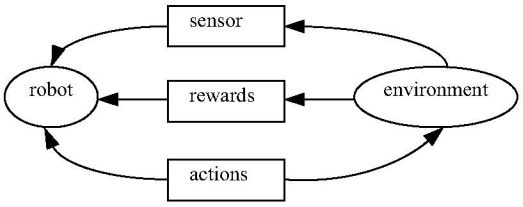

Fig. 3 Reinforcement learning model

distance informations provided by seven ultrasonic sensors. Fig.4 shows the sensors layout. The current interior state S_{int} are the moving velocity and the direction of the robot. So the static obstacle behavior is to learn the optimal mapping relationship $\pi : S(S_{ext}, S_{\text{int}}) \to A$.

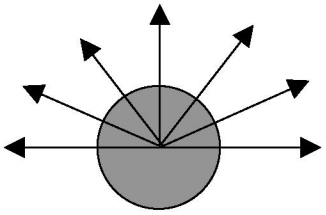

Fig. 4 Ultrasonic sensors layout

A. Reinforcement function design Equations

The learning task for mobile robot agent is described by the reinforcement function. An immediate payoff is given to a state-action pair. In order to reflex the evaluation of the robot performances better and more concretely, we adopt the sum of vectors pointing away from obstacles as the reinforcement function *r(t)* comparing with some references using constant as the reinforcement signals [9].

The reinforcement function of the static obstacles avoidance is designed as formula (1).

$$r(t) = \begin{cases} -1 & d_{bi} \le d_{\min} \\ -k\sum_{i=1}^{5} \frac{1}{d_i} & d_{\min} \le d_{bi} \le d_{\max} \\ 0 & d_{bi} > d_{\max} \end{cases} \qquad (1)$$

Where $d_{bi}(i=1\sim5)$ is the distance vectors pointing away from obstacles after normalization. *K* is the obstacle distance

parameter. Use it to restrict the value of *r(t)* between 0 and 1. The reinforcement function shows that more obstacles around the robot and nearer to the robot, little reward obtained. This means the robot is in a disadvantage environment. The value – 1 of *r(t)* explains that robot has collided with obstacle while 0 shows robot is in a open area. The value between 0 and 1 reflexes the obstacle avoidance degree. At last using if-then fuzzy control rules to describe formula (1).

B. Action selection Equations

The task of the mobile robot is to execute a series of actions, observe their consequence, then learn control strategy. The goal of the learning is to maximize the amount of reward it receives in the long term.

We suppose the environment is a discrete Morkov procedure of an finite states set. When the agent selects an action from the actions set, the environment accepts the action and executes it. Then the state changes to the next. The changing probability of the state is as follows.

$$prob[s = s_{t+1} / s_t, a_t] = P[s_t, a_t, s_{t+1}] \qquad (2)$$

Q-learning is an unsupervised learning method based on trial and error to find an optimal behavior strategy $\pi : S \to A$. The convergent theorem of Q-learning requires each state-action pair can be visited frequently. When we accumulate experiences and train examples, if the action is selected according to the random principle, some unnecessary pairs which have little Q value will occur frequently. This can bring the long learning time. While if we adopt ε greedy principle, the mobile robot agent will always select the action which has the maximal Q value. This can make some pairs which have little values will not be visited forever. Then we can not find the optimal control strategy. In this paper we adopt a novel prior Q search method (*PQA*) to rank the reward of each state-action pair according to the prediction values and accumulated experience. In the Q-learning each action is selected by probability. The state-action pair which has a high reward is given a high probability, but all state-action pairs probability are nonzero.

PQA predicts the moving direction β for the next time by solving the sum of vectors pointing away from obstacles based on potential field principle. Compare β with the direction of the executable actions to determine the prior rank of the actions for Q-learning to select an action. Here we illustrate the method briefly.

Use the angle α_{a_i} to denote the six actions of the action space A for mobile robot. They are $\alpha_{a_1} = 0^0$, $\alpha_{a_2} = -20^0$, $\alpha_{a_3} = -40^0$, $\alpha_{a_4} = +20^0$, $\alpha_{a_5} = +40^0$ and $\alpha_{a_6} = -180^0$. The sum of vectors pointing away from obstacles in the view of 180^0 in front of the robot is:

$$\vec{F} = \sum_{i=1}^{5} \vec{f}_i = F \angle \alpha \ (90^0 \le \alpha \le 270^0 \ or \ \alpha = 0^0) \qquad (3)$$

Four different operations are executed according to the value of α.

1) When $\alpha = 0^0$ select action α_{a_1}.

2) When $165^0 \le \alpha \le 195^0$ select action α_{a_6}.

3) When $90^0 \le \alpha < 165^0$ the prior rank of action space A is: $A = \{\alpha_{a_5}, \alpha_{a_4}, \alpha_{a_1}, \alpha_{a_2}, \alpha_{a_3}, \alpha_{a_6}\}$.

4) When $195^0 < \alpha \le 270^0$ the prior rank action space A is: $A = \{\alpha_{a_3}, \alpha_{a_2}, \alpha_{a_1}, \alpha_{a_4}, \alpha_{a_5}, \alpha_{a_6}\}$.

In the case of 3) and 4), the state-action pair which is ranked in the front is possible to be given a high reward. Then it can be selected by a high probability. So we can increase the learning speed evidently.

III. Q-Learining

The goal of the learning is to maximize the amount of reward it receives in the long term. The strategy of the Q-learning is to select the action which can have a maximal *Q(s,a)* value. It can be described as : $\pi^* = \arg\max_a Q(s,a)$. The key problem is to update Q values through interacting with the environment. The procedure of Q-learning is as follows:

1) Initialize $\hat{Q}(s,a)$ to be zero for all state s and action a.

2) Perceive the environment information and the interior state s_t of the robot, which $s_t \in S$.

3) Choose an action a_i according to the prior rank table based on the probability. The formula is:

$$p(a_i \mid s_t) = \frac{k^{\hat{Q}(s_t, a_i)}}{\sum_j k^{\hat{Q}(s_t, a_j)}} \qquad (4)$$

In which $p(a_i \mid s_t)$ is the probability when selecting the action a_i at state s_t. K is a constant and $k>0$. It determines the prior degree to select the $\hat{Q}$ which has a high value.*Paper*

4) Execute action a_i and arrive at the next state s_{t+1}, then gain the immediate reward r_{t+1}.

5) Update $\hat{Q}_n(s_t, a_i)$ according to Bellman formula.

$$\hat{Q}_n(s_t, a_i) = (1 - \alpha_n)\hat{Q}_{n-1}(s_t, a_i) + \alpha_n[r_{t+1} + \gamma \max_{a_i \in A}(s_{t+1}, a_i)] \qquad (5)$$

Where α_n is the learning rate, defined as:

$$\alpha_n = \frac{1}{1 + n(s_t, a_i)} \qquad (6)$$

Where $n(s_t, a_i)$ is the total times during the n times loop for the state-action pair (s_t, a_i). γ is the discount rate between 0 and 1. In this paper we use the value 0.9.

6) Return to *2)*

IV. Experiment And Discussion

The learning task for the mobile robot is to arrive at the goal avoiding the obstacles with the shortest steps (or in the shortest time). The start point, the goal and the obstacles are prescribed randomly in the training. One training step is an iterative procedure of the Q-learning. A training period includes N steps. The initial point is settled far away from the goal step by step in each new training period in order to cover more origination spaces. When the robot arrives at the goal, fails to fulfill the task or exceeds the maximal training steps, then end the training period and start a new one.

The training environment is a 600×400 pixel area. The obstacle densities of 10% and 20% are used. The figuration and the position of the obstacles are settled randomly. Select five start points in each environment and start training and learning in 200 groups respectively. Suppose the maximal steps are 5000 in one training period. Adopt *RA* and *PQA* strategies to select actions. Fig.5 and Fig.6 shows the experimental data and compares the result by the two strategies in different environment. Fig.7 draws the trajectory planned by *PQA* strategy in an environment of 10% obstacles densities.

In Fig.5 we can see that to fulfill task in a high obstacles densities environment need more average steps, namely a long learning time. In a same environment, the mobile robot requires shorter learning time by *PQA* method than by *RA* method. Fig.6 illustrates the completion rates of the task in the total training period. In the same way we can see *PQA* strategy has a higher completion rate in two different environments.

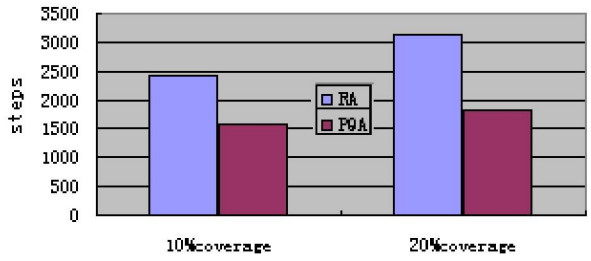

Fig. 5 Average steps of task completion in two different environments

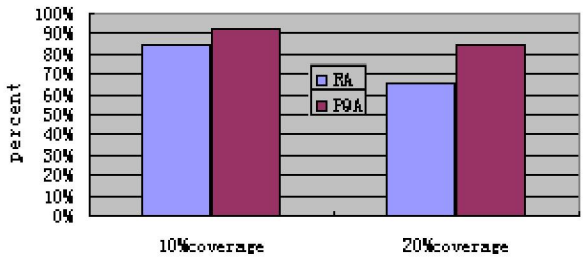

Fig. 6 Completion percent in two different environment

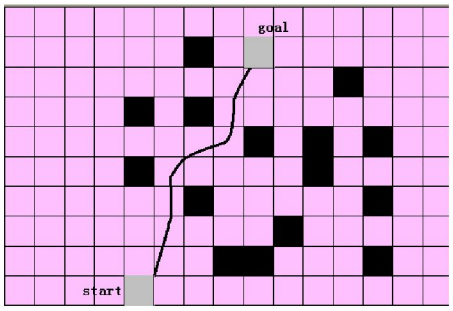

Fig. 7 Trajectory of the path planning by *PQA*

V. Conclusion

We can not have prior knowledge about the environment forever. Without prior knowledge, we can't deduce the appropriate controller parameter values. Furthermore obstacles densities and layout in the environment may be heterogeneous. Parameters that work well in one type of environment may not work well with another type. So it becomes necessary for the mobile robot to adapt on its own to what it finds. By adaptive learning, the controller constructs the direct mapping between the environment states and the actions. Then make the robot arrive at the goal with the shortest path avoiding the obstacles. In this paper we adopt the adaptive path planning based on the layered Q-learning method. In the learning of the static obstacle avoiding, *PQA* search strategy is adopted. Completion rates can be drastically improved by the method from the simulation result. Furthermore the method can be executed on-line. So it can become a possible and effective way to solve the path planning problem under dynamic and unknown environments.

Q learning, namely reinforcement learning, have some special problems besides the common difficulties share with other learning methods under uncertainty environment. In order to deal with this problems, in future, we will do some work through the following three means.

1) Develop the reinforcement learning method based on Markov procedure furthermore. Make the theory to meet with the requirement of the real application.

2) Simplify and deduce the complexities of the system according to the characteristics of the real task and environment. Increase the velocity of the convergence. This is the core problems of the reinforcement learning method.

3) Combine it with other methods, such as artificial neural network, evolutionary algorithm, to obtain better learning performance.

References

[1] Eduardo Zalama, Jaime Gomez, Mariano Paul, and Jose Ramon Peran, "Adaptive behavior navigation of a mobile robot," IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans, Vol.32, No.1, pp.160-169, January 2002.

[2] Song Qi, Han Jian-da, "UKF-based Active Modeling and M odel-reference Adaptive Control for Mobile Robots," Robot, Vol.27, No.3, pp.226-235, May 2005.

[3] R.A.Brooks, "Robust layered control systems for a mobile robot," IEEE Transactions on Robotics and Automation, vol.2, no.1, pp.14-23, 1986

[4] Dongbing Gu and Huosheng Hu, "Teaching Robots to Coordinate its Behaviours," Proceedings of the 2004 IEEE International Conference on Robotics & Automation, New Orieans, pp. 3721-3726 , April 2004.
[5] Naoyuki Kubota, "A Spikilng Neural Network for Behavior Learning of A Mobile Robot in A Dynamic Environment," 2004 IEEE International Conference on Systems, Man and Cybernetics, pp5783-5788, 2004
[6] Liu Zhao, Chen Jian-xu, "Design of soccer robot strategy base on AGA," JOURNAL OF HARBIN INSTITUTE OF TECHNOLOGY, Vol.37, No.7, pp.912-913, 2005.7.
[7] Panrasee Ritthipravat, Thavida Maneewarn, Djitt Laowattana and Jeremy Wyatt, "A Modified Approach to Fuzzy Q Learning for Mobile Robots, " 2004 IEEE International Conference on Systems, Man and Cybernetics, pp. 2350-2356, 2004.
[8] Brooks, "A Robust Layed Control System for a Mobile Robot, " IEEE Journal of Robotics and Autonomous, pp.1-10, 1986.
[9] Fan Bo, Pan Quan, Zhang hong-cai, "A Method for Multi-Agent Coordination Based on Distributed Reinforcement Learning," Computer simulation, Vol.22, No.6, pp.115-117, 2005.6.