

A Q-learning algorithm applied to the behavioural decision-making of affective virtual human

Yiwei ZHANG, Tianhuang CHEN

Computer Science and Technology

Wuhan University of Technology

No.1186, Heping Boulevard, Wuchang District, Wuhan, Hubei, CHINA

Email: yiwei2020@qq.com

Abstract—Traditional Q-Learning algorithm has problems of data transmission lag and its environmental reward model is too simple. It cannot be well applied to the reinforcement learning of affective virtual human behaviour decision. Analogizing the thought of human's self-reflection in this paper, a improved Q-learning algorithm is proposed, which can be easily applied in behavioural decision-making of affective virtual human. The Q-learning algorithm in this paper not only strengthens the behaviour strategy with better learning cycle and weakens the behaviour strategy with worse learning cycle by the way of self-reflection reward, but also picks up the speed of the effect of behavioural decision feedback to state-action pair in a learning cycle, thus improving the convergence rate of Q-learning algorithm in affective virtual human's behavioural decision-making. The algorithm aims at helping affective virtual human carry out path optimization in a two-dimensional grid environment in the simulation test. The results show that the improved Q-learning algorithm is significantly faster than the traditional Q-learning algorithm in achieving the optimal control strategy with an average of 43.7 learning cycles. The validity of the algorithm is verified.

Keywords—Reinforcement learning; Q-learning algorithm; affective virtual human; behavioural decision-making; environmental reward model; affective model

I. INTRODUCTION

The vigorous development of the current artificial intelligence technology is widely used in all aspects of people's lives, and it prompts the intelligent virtual human technology [1]. In order to meet the rigid requirements of human interaction, researchers are eager to develop a highly intelligent virtual human. However, the intelligence of lacking emotion cannot be called real intelligence. Therefore, it has become a more and more important research direction for intelligent virtual human to integrate artificial emotion in the field of artificial intelligence [2]. At present, the main research content of artificial emotion is emotion recognition, emotion modelling and emotion expression [3][4]. The study of artificial emotion should not be limited to this and the influence of emotion on the autonomous learning and behaviour decision of virtual human should be fully realized. The introduction of artificial emotion in the virtual human behaviour decision-making can make affective virtual human simulate human behaviour more vividly.

Reinforcement learning is one of the effective methods in machine learning [5]. The representation of the Q-learning algorithm is introduced into the affective virtual human behaviour decision. Affective virtual human with learning ability could make more humanized and friendly decision than general intelligent virtual human. First of all, affective virtual human will gain emotion reaction based on external environment, then choose an appropriate action according to current emotional state, ultimately reach the target state. The improved Q-Learning algorithm can strengthen the behaviour decision-making of better learning cycle, weaken the behaviour decision-making of worse learning cycle, and make the affective virtual human possess the ability of autonomous learning and behaviour decision.

II. EMOTION MODELING OF VIRTUAL HUMAN

Emotions are perceptual experience for human to view objective things. It can help people to choose the appropriate behaviour in different situations, and improve the adaptability of humans in the complex environment [6]. Studies show that human emotions play an irreplaceable role in autonomous learning and behaviour decision-making[7]. If there is no emotion, it may be a problem for people to decide what clothes should wear when get up. The amygdala, which is located in the limbic system of the cerebral cortex, is able to quickly acquire sensory inputs and then respond. Emotions can lead to the generation of behavioural actions without relying on rational thinking. This rapid response can help people to adapt to the external environment. This paper follows the mechanism of emotion itself triggering behaviour, and reach the function of emotion regulation and control of affective virtual human, and construct a virtual human emotion model based on event evaluation process [8]. Affective virtual human feels the external stimuli, encodes it in the form of event object, and then carries on the cognitive appraisal to the event object and transfer evaluation result to the emotion model to produce corresponding emotion state. The emotion output of the emotion model according with the human emotion change rule. Emotional model system is as shown in Figure 1.

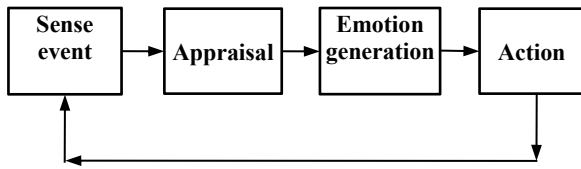


Figure 1. Emotional model system

At first, the external stimuli is encoded as event object e_t . Then the event object is evaluated from five-dimension (Desirability, Expectedness, Outcome Probability, Suddenness, Cause Harm) [9]. e_t is converted to a five dimensional cognitive evaluation vector $X_t(d_t, e_t, s_t, op_t, ch_t)$. Afterwards, $X_t(d_t, e_t, s_t, op_t, ch_t)$ is input into the emotional model as a parameter, and the output is emotional state. Affective virtual human will make appropriate behaviour decision according to the emotional state.

III. Q-LEARNING ALGORITHM

The basic content of reinforcement learning is to achieve the best mapping from state to action through finite trial-and-error learning in order to obtain the maximum reward function value [10]. The Q-learning algorithm is widely used in the search of the optimal behavioural control strategy for affective virtual human. It is an evolutionary form of the Markov decision process [11]. The goal of Q-Learning is to find a strategy π that maximizes the cumulative discount reward R_d^π obtained during the learning cycles:

$$R_d^\pi = \sum_{t=0}^{\infty} \gamma^t r_{t+i} \quad (1)$$

In the formula: $\gamma (0 < \gamma < 1)$ is a discount factor; t is a natural number, representing the learning cycle; r_t is immediate reward after implementation of each action; i is the number of time steps and it represent the cost of reaching the optimal strategy.

The Q-value is the mathematical expectation. Affective virtual human get the reward discount by choosing action a_t according to the state s_t .

$$Q^\pi(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} P(s_t, a_t, s_{t+1}) V^\pi(s_{t+1})$$

$$V^\pi(s_t) = \max_{a \in A} Q^\pi(s_t, a_t) \quad (2)$$

In the formula: $S = [s_1, s_2, \dots, s_t]$ is the state set of the affective virtual human in the environment; $A = [a_1, a_2, \dots, a_t]$ is the action set for the corresponding state; r_t is the immediate reward of action at under state s_t ; $P(s_t, a_t, s_{t+1})$ is the probability that the affective virtual

human performs the action a_t from the state s_t to the next state s_{t+1} .

The Q-Learning method is implemented by the following recurrence formula:

$$Q_{t+1}(s_{t+1}, a_{t+1}) = (1 - \lambda) * Q(s_t, a_t) + \lambda * [r_t + \gamma * \max Q(s_{t+1}, a_{t+1})] \quad (3)$$

In the formula: λ is learning rate for controlling learning speed. In a certain range of values, the greater the λ , the faster convergence, but λ may lead to non-convergence when λ is too large. When $t \rightarrow \infty$, the formula (3) carries on the infinite iteration, and the $Q_t(s, a)$ will converge to the optimal policy with the probability 1 [12].

IV. THE Q-LEARNING ALGORITHM APPLIED TO THE BEHAVIORAL DECISION-MAKING OF AFFECTIVE VIRTUAL HUMAN

Q-Learning goal is to achieve the optimal control strategy according to the reward signal in the unknown environment, which can be understood as an online optimal decision learning process. Q-Learning is a learning algorithm not relying on the environment model. At the beginning, the affective virtual human does not know the external environment information at all, and it can only learn by limited trial and error method [13]. In the traditional Q-Learning algorithm, it is assumed that there are three adjacent state S1, S2 and S3. If the state S1 wants to obtain the state S3's information feedback, it needs to go through 2 learning cycles.

The first cycle: $Q(s_2) \rightarrow Q(s_1) \rightarrow Q'(s_1)$

$$Q(s_3) \rightarrow Q(s_2) \rightarrow Q'(s_2)$$

The second cycle: $Q'(s_2) \rightarrow Q'(s_1) \rightarrow Q''(s_1)$

The traditional Q-learning algorithm is deduced that the data transfer has obvious hysteresis. Q-value needs to go through at least n times' feedback learning to get the Q-value from the n states. It needs to go through a lot of iterations of learning cycle to feedback to the current state. If the follow-up decision-making of affective virtual human is error, its iterations will be very large, which is an important reason to limit the convergence rate of traditional Q-Learning algorithm [14].

In general, it would be reasonable to reward those better learning cycles and punish those worse learning cycles appropriately. The reward model of the traditional Q-learning algorithm is relatively simple, and the reward for a single learning cycle to reach the target state is constant R, which cannot distinguish the quality of learning cycles.

Aiming at the data transfer lag and the problem that reward model of the traditional Q-learning algorithm are too simple; the Q-learning algorithm applied to the decision-making of affective virtual human is put forward. According to the initial emotional state S_1 , the affective virtual human chooses the

action a_1 , the execution of state-action pair (s_1, a_1) reaches another emotional state S_2 . Instead of executing equation (3) immediately, the state-action pair (s_1, a_1) is stored in the stack structure. Affective virtual human select the corresponding action a_2 according to the state S_2 , then execute (s_2, a_2) and put (s_2, a_2) into the stack. Continue until the target state S_n is reached. Variable L of learning cycle step size and historical minimum step size $L_{\min}$ are introduced. If the current learning cycle $L < L_{\min}$, then $L_{\min} = L$. Three kinds of Self-reflection reward r_1, r_2, r_3 of affective virtual human are introduced, which correspond to the best, the better and the worse learning cycle respectively. Assuming that R is a known reward to reach the target state and $r_1 > r_2 > r_3$, the reward function is:

$$\text{reward} = \begin{cases} R + r_1 & L < L_{\min} \\ R + r_2 & L = L_{\min} \\ R + r_3 & L > L_{\min} \end{cases} \quad (4)$$

After reaching the target state, according to the reward of current learning cycle, the improved Q-learning algorithm updates the Q-value of state-action pair stored previously in the stack.

The improved Q-learning algorithm enhances the state-action pairs of the better learning cycle in the form of self-reflection reward of affective virtual human, and speeds up the feedback influence of the learning cycle reward to the previous state-action pairs and improves the learning efficiency of the affective virtual human. The Q-learning algorithm applied to the behavioural decision-making of affective virtual human is described as follows:

Algorithm 1: The Q-learning algorithm applied to the behavioural decision-making of affective virtual human

- 1) Initialize mathematical expectation $Q(s_i, a_i)$ of each state-action pair, initialize the discrete event e_i of the emotion model;
- 2) Update $e_i(e_1, e_2, \dots, e_n)$ and convert it to a five-dimensional cognitive evaluation vector $X_i(d_i, e_i, s_i, op_i, ch_i)$;
- 3) The output value s_i of the emotion state is updated according to $X_i(d_i, e_i, s_i, op_i, ch_i) \rightarrow S_i$;
- 4) Observe s_i , if the target state is reached, skip to step 7, otherwise go to step 5;
- 5) The virtual human performs the action a_i selected by rules of action selection in the environment model;
- 6) Put state-action pair (s_i, a_i) into stack, jump to step 2;
- 7) Update the reward of reaching the target state according to the formula (4);

- 8) Pop state-action pairs (s_i, a_i) from the stack; Update Q-value of each state-action pair according to formula (3);
- 9) End.

V. EXPERIMENTAL SIMULATION OF THE IMPROVED Q-LEARNING ALGORITHM

A. Description of experimental simulation

Experimental environment is shown in Figure 2, the task of the affective virtual human is path optimization in the two-dimensional barrier-free 20×20 grid environment. In addition to the boundary position, the basic actions of the affective virtual human in the environment are: up, down, left and right. Each grid represents a state of the affective virtual human. In order to test the performance of the improved algorithm more accurately and conveniently, grid coordinates are substituted for the emotional states. The starting position of the affective virtual human is $s(0,0)$ and the target position is $s(17,18)$. All things in the environment are static. Initially, environment model is unknown for the affective virtual human.

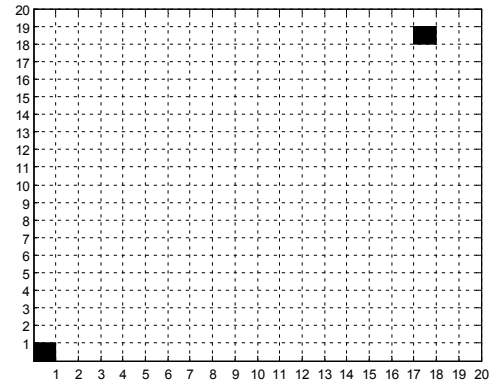


Figure 2. Experimental environment

Affective virtual human can obtain fixed reward $R = 50$ when completing the navigation task. If the current learning cycle is a better learning cycle ($L < L_{\min}$), then the self-reflection reward is $r_1 = 50$, if it is a poor learning cycle ($L > L_{\min}$), then $r_3 = -100$, if it is a general learning cycle ($L = L_{\min}$), then $r_2 = -30$. Discount factor $\gamma = 0.95$, learning efficiency $\lambda = 0.3$.

B. Simulation experiment result analysis

The affective virtual human uses the traditional Q-learning algorithm and the improved Q-learning algorithm respectively to complete the path optimization task in the two-dimensional barrier-free grid environment. The average values of 10 groups were obtained separately from the experimental data. The experimental results are shown in Figure 3.

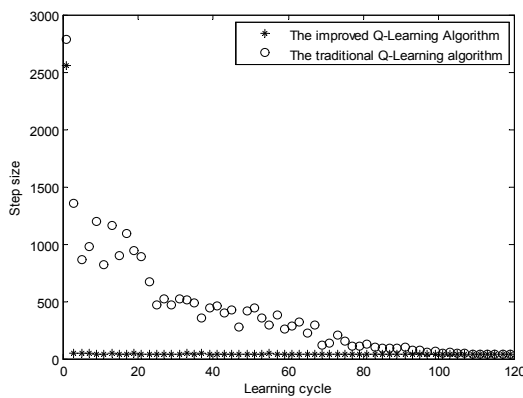


Figure 3. Experimental simulation results

In the first Learning cycle, the two Q-learning algorithms know nothing about the environment, and they all have to go through a long period of exploration to reach the target state. In the subsequent learning cycle, the performance of the improved Q-learning algorithm is far more superior than the traditional Q-learning algorithm. The reason is that in traditional Q-learning, data transfer between two states separated by n needs to go through multiple learning cycles and complete at least n feedbacks. However, the data transfer between n states can complete in a learning cycle in the improved Q-learning algorithm, thus greatly speeding up the data transfer efficiency. This feature is very similar to human learning habits that people tend to make decisions that have been successful before. At the same time, in order to avoid falling prematurely into the local optimum, the exploration factor is introduced into the selection strategy. After 120 learning cycles, both algorithms converge to the optimal strategy, but the former is significantly faster than the latter in achieving the optimal control strategy with an average of 43.7 learning cycles. The improved Q-learning algorithm makes a distinction of reward among the three kinds of learning outcomes in a learning cycle. Therefore, the improved Q-learning algorithm has less volatility than the traditional Q-learning algorithm in the process of gradual convergence.

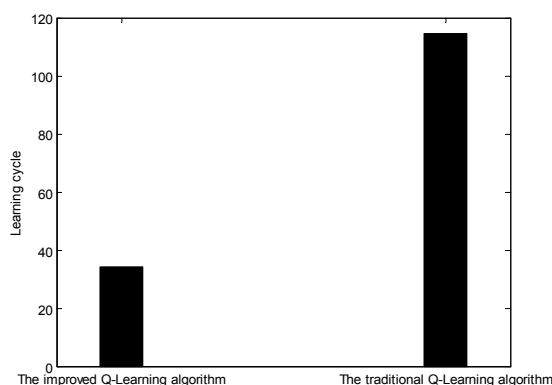


Figure 4. The average learning cycle for the first occurrence of the optimal step size

The average learning cycle for the first occurrence of the optimal step size is shown in Figure 4. Compared with the traditional Q-learning algorithm, the average number of learning cycles for the first occurrence of the optimal step of the improved Q-learning algorithm is reduced by 70.5. Traditional Q-learning algorithms only rely on updating the Q-value to carry out the exploration of the unknown environment, but it does not fully exploit more useful environmental information in the exploration process, while the improved Q-learning algorithm is obviously "smarter".

VI. CONCLUSIONS

By analogizing the thought of human's self-reflection, the paper proposes an improved Q-learning algorithm and it can be applied to the behaviour decision-making of affective virtual human. The improved Q-learning algorithm also knows nothing about the environment, but it can simply carry out summary and self-reflection in the explored environment. The experiment results show that the improved Q-learning algorithm can greatly improve the learning efficiency of the behaviour decision-making of affective virtual human, which validates the effectiveness of the algorithm. The future work will further verify the advancement and reliability of the algorithm in a larger state-action set and promote the practical application of the algorithm.

REFERENCES

- [1] Zhao Wei ,Xie Xiaofang.Development of Virtual Human Technology and Its Engineering Application[J].Journal of System Simulation,2009,21(17):5473-5476.
- [2] Wang Zhiliang,Wang Wei,Gu Xuejing,Zheng Siyi.Research Summarization of Humanoid Expression Robot with Emotion[J]. Computer Science, 2011, 38(1):34-39.
- [3] Zhu Yuhong,Wei Jinhai,Mao Junxin.Summary of Artificial Emotion[J].Journal of Jiangnan University(Natural Science Edition),2012,11(4):497-504.
- [4] Terada K,Yamauchi A,Ito A.Artificial emotion expression for a robot by dynamic color change[C] Paris France:IEEE Press,2012:314-321.
- [5] Chen Xue-song,Yang Yimin.Reinforcement learning: survey of recent work[J]. Application Research of Computers,2010,27(8):2834-2838.
- [6] Ashkanasy N M, Humphrey R H. Current Emotion Research in Organizational Behavior[J]. Emotion Review, 2011, 3(2):214-224.
- [7] Spence S. Descartes' Error: Emotion, Reason and the Human Brain[J]. Bmj Clinical Research, 1995, 310(6988):1213-1213.
- [8] Pan Zhigeng,Yang Hongwei,Liu Zhen.Overview of Emotion for Virtual Agent[J].Journal of Computer-Aided Design & Computer Graphics,2007,19(12):1509-1516.
- [9] Jain S,Asawa K. EMIA:Emotion Model for Intelligent Agent[J].Journal of Intelligent Systems,2015,24(4):449-465.
- [10] Yan Youbiao,Chen Yuanyan.A Survey on Machine Learning and Its Main Strategy[J].Application Research of Computers, 2004, 21(7):4-10.
- [11] Yang Ping,Bi Yiming,Liu Weidong. Decision-making model of tactics maneuver agent based on fuzzy Markov decision theory[J].Systems Engineering and Electronics,2008,30(3):511-514.
- [12] Watkins C J C H, Dayan P. Q -learning[J]. Machine Learning, 1992, 8(3):279-292.
- [13] Yu Dongmei,Han Xiaoxin,Li Ding,Xia Min. Research on Computable Model of emotional Interaction Based on Q-learning Algorithm[J].Computer Engineering,2012,38(10):277-279.
- [14] Xu Ya.Research on Path Planning for Mobile Robot Based on Reinforcement Learning[D].Jinan:Shandong University,2013.



Yiwei Zhang, born in Henan Province, China, September 20, 1991. A bachelor's degree in computer science and technology. Graduate student in school, the main research areas are: Affective computing and reinforcement learning.



Tianhuang Chen, born in Hubei Province, China, Professor in school of Computer Science and Technology, Wuhan University of Technology. The main research areas are: affective computing, database technology.