Learning Proxemic Behavior Using Reinforcement Learning with Cognitive Agents

Cristian Millán-Arias[1], Bruno Fernandes[1], Francisco Cruz[2,3]
[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife-PE, Brazil
[2]School of Information Technology, Deakin University, Geelong, Australia
[3]Escuela de Ingeniería, Universidad Central de Chile, Santiago, Chile
Emails: {ccma, bjtf}@ecomp.poli.br, francisco.cruz@deakin.edu.au

Abstract—**Proxemics is a branch of non-verbal communication concerned with studying the spatial behavior of people and animals. This behavior is an essential part of the communication process due to delimit the acceptable distance to interact with another being. With increasing research on human-agent interaction, new alternatives are needed that allow optimal communication, avoiding agents feeling uncomfortable. Several works consider proxemic behavior with cognitive agents, where human-robot interaction techniques and machine learning are implemented. However, environments consider fixed personal space and that the agent previously knows it. In this work, we aim to study how agents behave in environments based on proxemic behavior, and propose a modified gridworld to that aim. This environment considers an issuer with proxemic behavior that provides a disagreement signal to the agent. Our results show that the learning agent can identify the proxemic space when the issuer gives feedback about agent performance.**

Index Terms—**Cognitive Agents, Proxemics, Reinforcement Learning**

I. INTRODUCTION

Proxemics is the study of spatial behavior, concerned with territoriality, interpersonal distance, spatial arrangements, crowding, and others aspects of the physical environment that affect behavior. The term was coined by Hall et al. [1], when he proposes a fixed measure of personal space, a set of regions around a person to delimit the acceptable distance to interact with other people. In recent years, human-agent interaction has taken hold in the scientific community. Furthermore, the humanization of agents is an expected event, given technological advances and human nature. Thus, an optimal interaction is necessary on both the agent and the person [2].

Reinforcement Learning (RL) is a learning paradigm that tries to solve the problem of an agent interacting with the environment to learn the desired task autonomously [3]. The agent must sense a state from the environment and take actions that affect it to reach a new state. The agent receives a reward signal from the environment that it tries to maximize throughout the learning for each action taken. The agent takes actions from its own experience, or can be guided by an external trainer that provides feedback [4], [5]. Proxemic behavior has been used in different areas with cognitive agents. For example, in human-robot interaction, it has been studied how people behave in the presence of an artificial agent or robot and how their perception of the personal space of the agent has been [6], [7]. Moreover, in machine learning, it has been studied how artificial agents sense the personal space of other cognitive agents, and it has been to identify and learn personal space [8], [9].

II. PROXEMIC BEHAVIOR IN COGNITIVE AGENTS

We propose a modified version of the gridworld problem. In this environment, an issuer is placed on one fixed state and is responsible for giving a signal of disagreement when the learning agent is too close. Two regions are defined around the issuer, the uncomfortable region and the target region. The uncomfortable region is those states that make up a square around the issuer. In this work, the uncomfortable region is only an area of negative reward. Similarly, the target region is those states that make up a square around the uncomfortable region. A new action, PING, is added to the four traditional ones (UP, DOWN, LEFT, RIGHT). This action represents a communication signal with the issuer, i.e., the agent send a ping to ask the issuer if it is in the target region. The task finishes in three conditions:

- *Condition C1*: When the agent reaches the issuer.
- *Condition C2*: When the agent sends ping five times out of the target region.
- *Condition C3*: When the agent sends ping in the target region.

The reward function is defined as:

$$\rho = \rho_{issuer} + \rho_{grid}, \qquad (1)$$

where ρ_{issuer} is a numerical reward given by the issuer when the agent performs the PING action, and ρ_{grid} is the environment reward defined as:

$$\rho_{grid} = \begin{cases} -1.0 & \text{if} \quad \text{conditions C1 or C2} \\ -0.8 & \text{if} \quad \text{reach to uncomfortable region} \\ -0.4 & \text{if} \quad \text{give incorrectly ping} \\ -0.1 & \text{if} \quad \text{another state} \\ +1.0 & \text{if} \quad \text{condition C3} \end{cases} . \quad (2)$$

In our experiments, we use a 10×12 gridworld, the issuer is placed in $(6, 8)$ and maintains fixed during the training. Each agent starts the training in the superior left corner of the grid $(0, 0)$.

III. EXPERIMENTAL RESULTS

To explore the behavior of agents in a proxemic environment, we apply the Q-Learning algorithm [10] in the modified gridworld problem. We use ϵ-greedy to select random actions. We set the values of ϵ and learning rate α at 0.6 and the discount factor γ at 0.9. In our experiments, 100 agents are trained with 10000 time-steps.
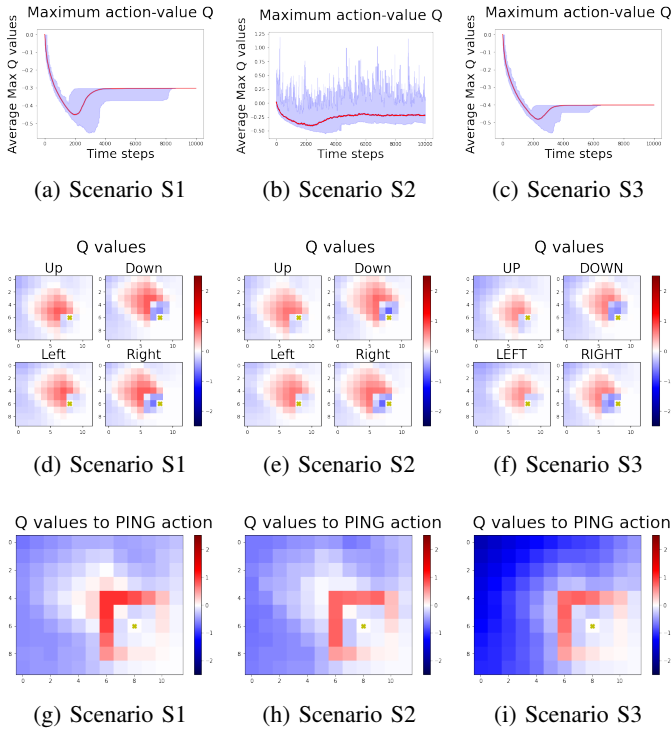
Fig. 1: Results of our experiments in three scenarios. The first row (1a-1c) shows the average of maximum Q values per time step, the shaded area is between the maximum and minimum Q values from 100 agents. The second row (1d)-1f) shows the average of final Q value for the actions UP, DOWN, LEFT, and RIGHT. The last row (1g-1i) show the average final Q values for the PING action.

In our experiments, we consider three scenarios:

- Scenario S1: When $\rho_{issuer} = 0$ in each PING action selection.
- Scenario S2: When ρ_{issuer} is a random value in $(-1, 1)$.
- Scenario S3: When $\rho_{issuer} = -\dfrac{d_1(P_{agent}, P_{issuer})}{d_1(P_{start}, P_{issuer})}$, where $d_1(\cdot, \cdot)$ is the L_1 distance, P_{agent}, and P_{issuer}, and P_{start} are the position on the grid from the agent, the issuer and the start point. This value indicates that states farthest from the issuer have a low reward, and the denominator standardizes it on (-1, 0).

Fig. 1 shows the results of our experiments in the three scenarios. In terms of maximum Q-value per time step (Fig. 1a-1c), there exists more variability when the issuer gives random rewards. However, according to the other scenarios, the maximum Q-value reaches values around 1 in several time steps. On the other hand, when the issuer gives rewards based on distance, the variability of Q-values is minor.

The Q values for each action are more focused on the target region than in the other scenarios when the issuer gives rewards based on distance (Fig. 1d-1f). This is expected due to the agent receives additional information on how to move through the grid.

Concerning the PING, the issuer reward highlights the importance of giving ping in the target region, as shown in

Fig. 1g-1i. The graph shows how the states of the target region have only Q-values greater than zero. While with the lack of information (Scenario S1), states outside the target region have Q-values greater than zero.

IV. CONCLUSIONS AND FUTURE ALIGNMENTS

In this paper, we study the agent performance in an environment based on proxemic behavior. We implement a modified gridworld problem, where an issuer agent gives a signal of disagreement when the RL agent performs a ping action. Moreover, the target is a proxemic region around the issuer instead of the traditional goal state. In our experiments, we consider three scenarios based on the reward type that the issuer gives. Our results show that the agent can reach the target region, even when the issuer does not give information. On the other hand, the issuer reward gives more information about performing the ping action, even when random information is given. Thus, the agent can identify the correct region to give a ping signal.

In our implementation, we consider that the proxemic region is symmetric around the user. Thus, in our future alignments, we intend to implement asymmetric proxemic regions to mimic the human proxemic behavior. Also, we contemplate implementing no fixed proxemic region, considering that the proxemic space changes by external factors. Finally, we intend to study the agent performance in more complex environments, involving other algorithms and techniques of reinforcement learning and deep learning.

REFERENCES

[1] E. T. Hall, R. L. Birdwhistell, B. Bock, P. Bohannan, A. R. Diebold Jr, M. Durbin, M. S. Edmonson, J. Fischer, D. Hymes, S. T. Kimball *et al.*, "Proxemics [and comments and replies]," *Current anthropology*, vol. 9, no. 2/3, pp. 83–108, 1968.

[2] N. Churamani, F. Cruz, S. Griffiths, and P. Barros, "icub: learning emotion expressions using human reward," *arXiv preprint arXiv:2003.13483*, 2020.

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: MIT press, 2018.

[4] C. Millán, B. Fernandes, and F. Cruz, "Human feedback in continuous actor-critic reinforcement learning," in *Proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, no. April, Bruges (Belgium), 2019, pp. 661–666.

[5] C. Millan-Arias, B. Fernandes, F. Cruz, R. Dazeley, and S. Fernandes, "A robust approach for continuous interactive actor-critic algorithms," *IEEE Access*, vol. 9, pp. 104 242–104 260, 2021.

[6] J. Mumm and B. Mutlu, "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 331–338.

[7] G. Eresha, M. Häring, B. Endrass, E. André, and M. Obaid, "Investigating the influence of culture on proxemic behaviors for humanoid robots," in *2013 IEEE Ro-Man*. IEEE, 2013, pp. 430–435.

[8] P. Patompak, S. Jeong, I. Nilkhamhang, and N. Y. Chong, "Learning proxemics for personalized human–robot social interaction," *International Journal of Social Robotics*, vol. 12, no. 1, pp. 267–280, 2020.

[9] N. Mitsunaga, C. Smith, T. Kanda, H. Ishiguro, and N. Hagita, "Robot behavior adaptation for human-robot interaction based on policy gradient reinforcement learning," *Journal of the Robotics Society of Japan*, vol. 24, no. 7, pp. 820–829, 2006.

[10] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.