Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 20(3): 181-188, Sep. 2022 Regular paper

Autonomous and Asynchronous Triggered Agent Exploratory Path-planning Via a Terrain Clutter-index using Reinforcement Learning

Min-Suk Kim[1], and Hwankuk Kim[2]*

[1]Department of Human Intelligence and Robot Engineering, Sangmyung University, Cheonan 31066, Korea
[2]Department of Information Security Engineering, Sangmyung University, Cheonan 31066, Korea

Abstract

An intelligent distributed multi-agent system (IDMS) using reinforcement learning (RL) is a challenging and intricate problem in which single or multiple agent(s) aim to achieve their specific goals (sub-goal and final goal), where they move their states in a complex and cluttered environment. The environment provided by the IDMS provides a cumulative optimal reward for each action based on the policy of the learning process. Most actions involve interacting with a given IDMS environment; therefore, it can provide the following elements: a starting agent state, multiple obstacles, agent goals, and a cluttered index. The reward in the environment is also reflected by RL-based agents, in which agents can move randomly or intelligently to reach their respective goals, to improve the agent learning performance. We extend different cases of intelligent multi-agent systems from our previous works: (a) a proposed environment-clutter-based-index for agent sub-goal selection and analysis of its effect, and (b) a newly proposed RL reward scheme based on the environmental clutter-index to identify and analyze the prerequisites and conditions for improving the overall system.

Index Terms: Intelligent Distributed Multi-Agent System (IDMS), Reinforcement Learning (RL), Sub-Goal. Environment-Clutter-Index

I. INTRODUCTION

An intelligent multi-agent distributed system (IMDS) is a monitoring system that achieves an agent's tasks in a geographically and computationally distributed environment. An IMDS has multiple agents and common or conflicting tasks for agent path planning [1,2]. It can provide flexibility and extensibility with some of the learned data for monitoring applications [3]. Such an intelligent system generally adopts the use of each of the agent-learning processes for autonomous path planning towards its respective goals (destinations). The system also requires the development of a computational multi-agent learning process in a large cluttered environment, where the agents have limited capabilities for path planning, and only have access to partially local information (knowledge) of their environment depending on the distributed computing node [3,4]. In such a large, cluttered environment, an agent can move randomly towards its goal. It is desirable for the agent to be intelligently equipped to move and avoid obstacles in the environment, and to be able to autonomously learn the shortest path planning to collect environmental knowledge in a minimum amount of time and steps [1,5].

IMDS is also provided by reinforcement-learning-based

Received 31 May 2022, **Revised** 11 August 2022, **Accepted** 17 August 2022
***Corresponding Author** Hwankuk Kim (E-mail: rinyfeel@smu.ac.kr, Tel: +82-41-550-5101)
Department of Information Security Engineering, Sangmyung University, Cheonan 31066, Korea

Open Access https://doi.org/10.56977/jicce.2022.20.3.181 print ISSN: 2234-8255 online ISSN: 2234-8883

autonomous multi-agent path planning for monitoring the environmental infrastructure and resources in the agent computational learning process [6]. Agent path planning based on reinforcement learning (RL) with multi-agent exploration enables the evaluation of a cumulative reward for every action and step. Optimized knowledge for the next available action is repeatedly needed by the learning process according to a learning policy [1,7]. RL is a machine learning algorithm, which is normally based on a reward provided by the environment in a state transition for an agent learning process [8-11]. The learning algorithm requires a system capable of autonomous acquisition and incorporation of knowledge [9]. It continuously improves and becomes more efficient as the learning process from an agent's exploratory experience to optimize an agent's learning performance in a time-varying environment [12,13]. Intelligent machine learning methods also require the study of computer algorithms to automatically improve the agent learning performance of path planning [14]. This scheme can attempt to determine a policy and learn a maximizing cumulative reward for a faster optimal path [15,16]. RL is typically used in multi-agent-based monitoring systems to solve the problem of learning strategies using an autonomous agent [7,17,18]. It has emerged as an area of memory capacity and computational power since the start of the use of learning algorithms [19] in multi-agent systems. RL is an intelligent learning scheme for dynamic environments with complex challenges, such as slow learning speed and hardware limitations. However, it has a problem that has been diminished owing to ongoing hardware improvements. The proposed method has been used to develop time-varying [12,20] and real-time applications [21] such as mobile robotics. The RL-based multi-agent approach can counter many different problems, such as machine learning, to solve multi-agent coordination and collaboration [22,23].

A. Contributions and Objectives

In this study, we focus on a multi-agent system with collaborative agents based on the proposed schemes using reinforcement-learning-based agent path planning. The scheme adopts and extends prior studies [1,7,24] to demonstrate the agent learning process in a distributed environment with one or more sub-goals, where multiple agents have different final goals (destinations) for agent path planning. In prior studies, RL was mainly used as the agent learning process to self-improve learning performance [7,10,14,23]. RL is also the study of machine learning algorithms to automatically attempt and find maximizing cumulative rewards for faster optimal path planning in terms of value and policy networks. The scheme is based on a sharing-information scheme, which is a communication scheme for an agent learning process [1,2,7]. Single and multi-agents can share their path exploration information with other agents supervised by the RL-based reward learning method on the existing local memory node. Without the intelligent scheme, the agent does not have the capabilities and resources in an entire given large terrain for learning performance because the initial random exploration becomes challenging and relatively nonconvergent. However, collaborative agent path planning, depending on the scheme of sharing information, can improve learning performance in a given large terrain. In addition, a new sub-goal-based RL reward function in an environmental clutter-index is proposed to improve the agent learning performance. In particular, the contributions of our research are (a) the approach of agent sub-goal selection to reduce smaller agent steps toward the given goal, where these agents can geographically explore the given environment, and (b) the approach as part of a newly proposed reinforcement-learning-based reward scheme for the autonomous and asynchronous triggering of agent exploratory phases.

B. Organization

Section 2 describes the system architecture. Sections 3 and 4 present our proposed method and the experimental results obtained using the proposed method, respectively. Section 5 presents conclusions and future work.

II. SYSTEM MODEL AND METHODS

The overall architecture shown in Fig. 1 is a clutter-index-based scheme based on a hybrid P2P [6,10]. The global environment (master node) initially assigns collaborating and monitoring agents in the terrain, which are distributed to the slave nodes geographically and computationally. Each collaborating agent that attempts exploration takes trials for synchronized real-time situational understanding [1]. The agents have awareness and decision-making to achieve their sub-goals and/or final goal (destination) via distributed reinforcement reward-based learning. The approach based on an agent sub-goal selection scheme in a multi-agent skill has been adopted from previous studies [1,7].

The multi-agent has limited resources and incomplete knowledge regarding when an agent performs exploration to find its goals (sub-goal and final goal) in a distributed environment. The agent lacks the capabilities and resources required to span the large terrain of its environment [1,15]. All agents distributed in a given large terrain, however, have capabilities to share the needed information over a network [16]. A given agent that surrounds neighboring agents does not need to run on the same exploring node. Each agent also does not have prior knowledge of the nodes on which other agents are running when communicating [1,25-27].

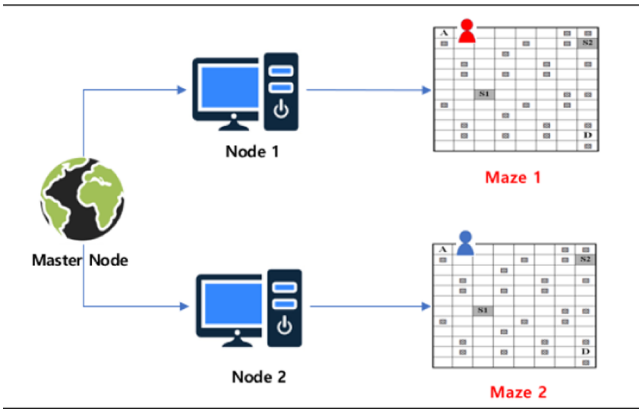

Fig. 1. Overall Hybrid Architecture (P2P / Master-Slave)

A. System Process and Structure

Fig. 2 shows the overall system diagram for a clutter-index-based technique using agent sub-goal selection [28, 29] in clutter-index multi-agent and goal system. Agents involved with sub-goal(s) are selected by the global environment for using a clutter-index scheme.

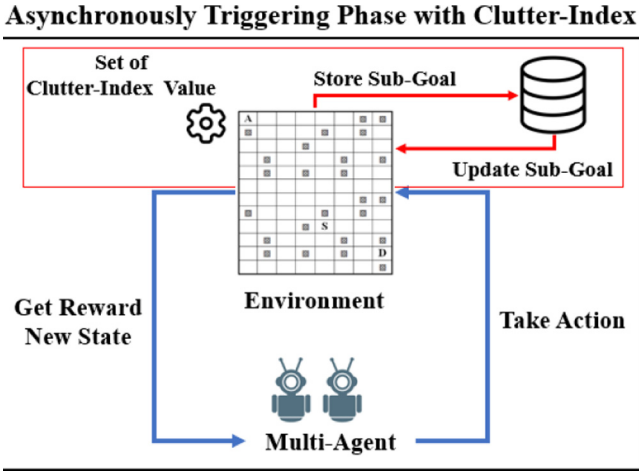

Fig. 2. Diagram of Clutter-Index-Based with RL-Based Reward scheme

This scheme can automatically and asynchronously trigger or switch between agent exploratory trials. In particular, each agent should start its exploratory path planning asynchronously and immediately after the multi-agent finishes the exploratory trials toward their sub-goals for overall agent learning performance.

B. Environment-Clutter-Index Node

There is a large terrain with different positioning resources and obstacles, according to their computationally coordinated environment. Fig. 3 shows two different sample mazes (8 × 12) that have a single or multiple sub-goals in the case of computing nodes. An agent can move to reach its goal (destination) in a computing environment via its sub-goal (s) to each portion. ⊠ denotes obstacles, A denotes an agent, S1 and S2 denote sub-goals, and D denotes the final goal (destination).

Fig. 4 shows agent path planning, where the agent can move in only 4-neighbor other directions, namely up, down, left, and right. In this case, a cluttered index value can be transferred to the master node to share the information with any other agent in the task for the next exploration trials. Agent (A) can be enclosed by obstacles and boundaries. In the case scenario shown in Fig. 4, the agent can also move in the downward direction to move to the next available position

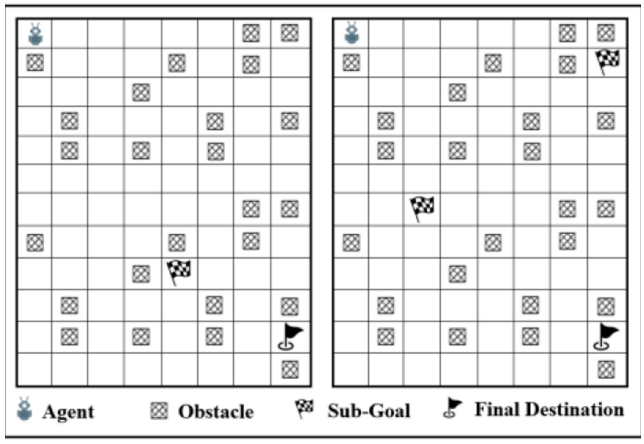

Fig. 3. Initial Real Terrain for Clutter-Index Value with Agent, Obstacle, Sub-Goal, and Final Destination

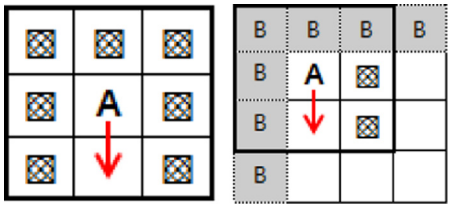

Fig. 4. Case of Scenarios with Obstacles, Boundaries, and Direction in case of Obstacle and Border

The proposed scheme using clutter-index value, which is described in Figs. 5 and 6, requires padding of the agent environmental terrain with obstacles and boundaries. The scheme has the clutter-index depending on the obstacles. The obstacles exist in the vicinity and neighborhood of an agent's position on the terrain. In this scheme, an agent does not distinguish between obstacles and padded boundaries. As shown in Fig. 5, a clutter-index value in scenario A (four directions) is derived from several obstacles and boundaries with an

enclosed position in an environment. Each clutter-index value is defined as four possible directions (up, down, right, and left) to move the next available position by an agent.

Otherwise, if an agent could take eight possible directions (above four directions and Up-right, Up-left, Down-right, and Down-left) to move the next available, a clutter-index value would be determined by eight possible directions, as in scenario B in Fig. 5. In the proposed method, we generally use scenario A for the agent learning process.

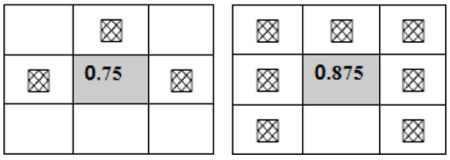

Fig. 5. Case of Scenario in Clutter-Index Values with Possible Directions (Left: A, Right: B)

C. Equation and Scheme

Here, a new policy equation is reflected by a clutter-index-based reward. It is expanded from the previous study [1,10], and the equation proposed with the reinforcement learning scheme is given below:

$$f(x,y) = O(x,y) - \frac{\varepsilon^2}{2}(x_g - x_c)[(x_g - x_c)^2 + (y_g - y_c)^2]^{-1/2} + \frac{\varepsilon^2}{visited} - \frac{\varepsilon^2}{reward} - \frac{\varepsilon^2}{index} \quad (1)$$

The equation is used to determine the next position when an agent moves in its environment. It has five terms, as follows. The first is the repelling term derived from obstacles that are found by agent path planning. The second term is the term attracting agents to their given destinations. The third term denotes the visited frequency (used in learning), and the fourth is a reward assigned by the global environment. The last term can help boost the agent path planning with environment-clutter-index values. As part of the evaluation, the function is defined by the number of obstacles and boundaries as well as the total number of available free directions in which an agent can move. The agent policy is to be learned as a function of agent positions in the computing node. In the proposed new scheme based on reinforcement-learning-based reward with clutter-index-based values, multiple agents can explore the next position to reach their final goal (destination) using the proposed clutter-index-based scheme.

Fig. 6 describes three agent learning steps; (1) creating a clutter-index table, (2) placing index values in the table during agent path planning in the first random trial, and (3) updating its clutter-index values when the agent explores its environment to reach its final goal (destination) via its single or multiple sub-goal (s).

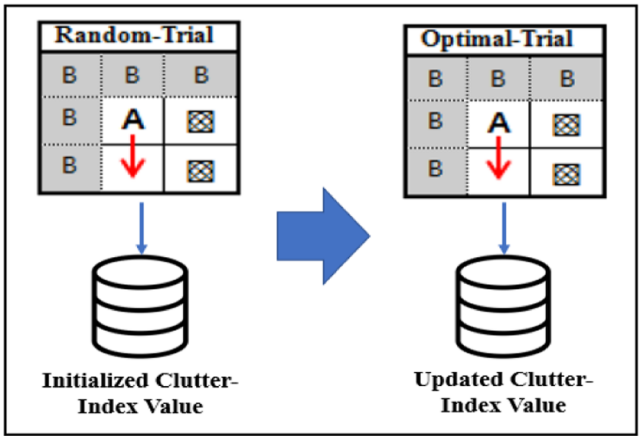

Fig. 6. Flow Chart of Agent Exploratory Trials Overview. (Note: Steps of the environmental clutter-index value)

III. RESULTS

A. Case of One Single Sub-Goal based on Clutter-Index Scheme

Here, we set the terrain and obstacle positions as experimental factors. The two input mazes had identical placement of resources and obstacles. Only one agent is used to try to find a single sub-goal; then, it can start exploring to find its final goal (destination). As shown in Fig. 7, the clutter-index table created by an agent's initial random exploration has different learned clutter-index values in a complex and cluttered environment. The clutter-index values can be evaluated while the agent (A) explores to reach its sub-goal (S). The

A	0	0.25	0				
	0.25	0	0				
0.5	0	0.25					
0.5		0.25	0.25	0			
0.25				0.25			
				0	0		
	0	0	0	0	0		
	0.25	0	0.5		0.25		0.75
0.5	0.25	0		S	0.25	0.25	0.5
0.5		0.25	0	0.25			
0.5				0.5			D
0.5							

Fig. 7. Clutter-index Values Discovered during the Agent (A) Path-Planning in the First Random-Exploration Trial to the Sub-Goal (S)

status of the table needs to be continuously updated during the exploratory trials until the agent reaches its sub-goal and final goal (destination).

As shown in Fig. 8, agent (A) is already discovered/ defined with clutter-index values from the first random trial. After the agent random exploration, the grey positions in the table are new clutter-index values discovered/updated towards its sub-goal. Although the agent learns some clutter-index values in the first random trial, the index values should be continuously defined and optimized by the agent learning process until the agent path planning reaches its final goal (destination).

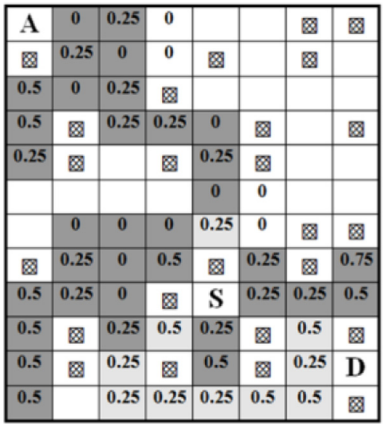

Fig. 8. Clutter-Index Values Discovered during the Agent (A) Path-Planning in the Exploratory Trials to the Sub-Goal and Final Goal (Destination).

B. Case of Two Different Sub-Goals based on Clutter-Index Scheme

As shown in Figs. 9 and 10, we set the different scenarios as follows: A denotes an agent, S1 and S2 denote two sub-goals, and D denotes the final goal (destination). S1 and S2 are controlled by the same computing node, but in different positions. The sub-goals (S1 and S2) have different clutter-index values; as such, S1 has a free index value (0), where there is no obstacle or boundary surrounding the sub-goal, while S2 has a cluttered index value (0.75), where there are two obstacles and one boundary to enclose the sub-goal. The learning performance of an agent differs depending on how the clutter-index value in the environment is optimized. The clutter-index value tables are learned by an agent (A) to the destination (D) via the sub-goals (S1 and S2). The agent (A) discovers/defines the clutter-index values in its first random trial; then, it can continuously try to discover/update the clutter-index values to find the best-optimized values during its exploratory trials towards its goal(s) (destination). It correctly selects the first sub-goal (S1) with the overall learning performance because S1 has a lower clutter-index value. Fig. 10 shows that the agent selects another sub-goal (S2) after its intelligent exploratory path planning. The clutter-index values are discovered/updated by the agent learning exploration until the goal is reached.

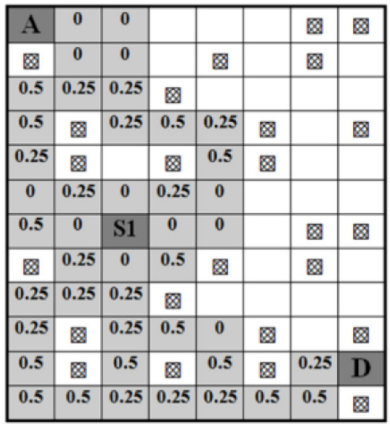

Fig. 9. Clutter-Index Values Discovered during the Agent (A) Path-Planning in the First Random-Exploration Trial to the Sub-Goal (S1).

According to the experiments, we present and compare the experimental results with different sub-goals for agent learning performance. The agent runs to migrate one node in search of its sub-goal(s) or final goal (destination) to the different positions of an obstacle, initially unknown to the agent. Some of the positions of obstacles are discovered or collected into environmental knowledge during the agent learning exploratory trials using a new clutter-index value scheme toward the goals.

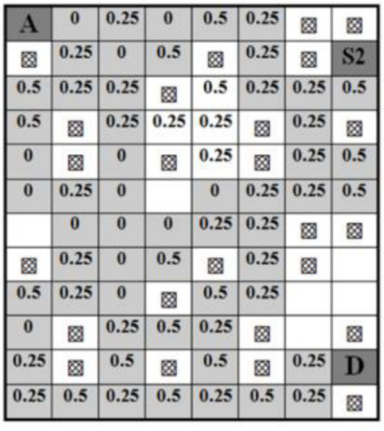

Fig. 10. Clutter-Index Values Discovered during the Agent (A) Path-Planning in the First Random-Exploration Trial to the Sub-Goal (S2).

C. Reward-based Clutter-Index Scheme: Single Sub-Goal

Fig. 11. shows the relationship between the total number of steps versus the total number of trials that an agent takes

to reach its final goal (destination) via one sub-goal, while using agent path planning with the proposed clutter-index-based reward scheme based on RL. In other words, the agent can take a smaller number of steps to reach its final goal (destination) with the proposed clutter-index-based reward scheme relative to the case where a clutter-index-based reward scheme is used.

It also uses automatic and asynchronous triggered agent exploratory trials. The results in Fig. 11 show the different plots of the effect of the environmental clutter-index in the case of an agent automatically and asynchronously triggering an exploratory phase to improve agent learning performance. These figures also show the number of trials required by the agent learning process to reach the final goal (destination) via a single sub-goal.

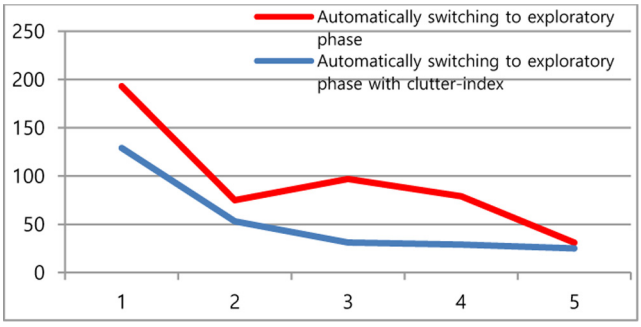

Fig. 11. Steps (Y) vs Trials (X) - Agent, Sub-goal, and Final Goal (Destination) (Note: With Clutter-Index-based vs Without Clutter-Index-based).

D. Reward-based Clutter-Index Scheme: Different Sub-Goals

There are additional experimental results for the proposed clutter-index-based reinforcement reward scheme via two selected sub-goals, with a comparison of the agent learning performance. In this scenario, agents can select each sub-goal in a complex and cluttered part of the environment with a proposed clutter-index-based reward scheme. As shown in Fig. 12, the two different sub-goals considered in the case scenario are located at the same node but at different positions in the terrain. The two sub-goals (S1 and S2) have different clutter-index values; S1 has a free index value (0) and S2 has a clutter-index value (0.75). Fig. 12 shows that each sub-goal (S1 or S2) provides the minimum number of steps in the agent exploratory trials to reach its final goal (destination) in comparison with using RL with a clutter-index-based reward scheme. Therefore, the agent finally chooses sub-goal (S1) for the minimum number of exploratory steps and better results for the agent learning process and system performance.

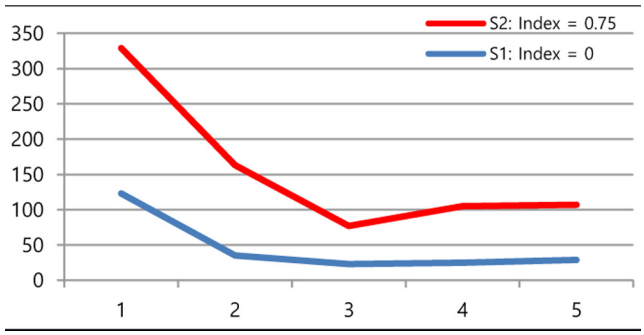

Fig. 12. Steps (Y) vs Trials (X) - Agent, Sub-Goal (S1, S2), and Final Goal (Destination) (Note: Agent has two different sub-goals, S1 has clutter-index value (0) and S2 has clutter-index value (0.75)).

IV. DISCUSSION AND CONCLUSIONS

The contributions of this research are the analysis and development of a multi-agent architecture performance and a new reinforcement learning reward, where an agent deduces the next position to reach its final destination using a clutter-index-based scheme on the overall system performance and the agent learning process through path planning. Each of the agents begins its exploratory trials asynchronously by following the agent learning steps: Step A: globally initialize and create a clutter-index value table; Step B: Randomly place all index values while exploring agent path planning; Step C: Discover and update the index values from the second exploratory trial until the agent reaches its final goal (destination) via the sub-goal (s). The experimental result shows that an agent should take an environment clutter-index using sub-goal selection to minimize the total number of learning steps. The agent also needs to select a new agent environment clutter-index-based scheme with RL-based reward for improving the agent learning performance.

Future work includes further analysis of optimized sub-goal selection in more realistic directions to eventually improve the self-play multi-agent learning scheme of high performance, followed by collaboration and competition in the intelligent multi-agent learning process. The work will additionally study the acquisition case of bad knowledge, with single or multiple agents acquiring and inheriting non-useful knowledge to collaborate with other agents in a distributed multi-agent environment.

ACKNOWLEDGEMENTS

This research was funded by a 2021 research grant from Sangmyung University.

REFERENCES

[1] M. -S. Kim, "A study of collaborative and distributed multi-agent path-planning using reinforcement learning," *Journal of The Korea Society of Computer and Information*, vol. 26, no. 3, pp. 9-17, Mar. 2021. DOI: 10.9708/jksci.2021.26.03.009.

[2] D. B. Megherbi, M. Kim, and M. Madera, "A study of collaborative distributed multi-goal and multi-agent based systems for large critical key infrastructures and resources (CKIR) dynamic monitoring and surveillance," in *IEEE International Conference on Technologies for Homeland Security*, Waltham: MA, USA, pp. 687-692, 2013. DOI: 10.1109/THS.2013.6699087.

[3] Y. Bicen and F. Aras, "Intelligent condition monitoring platform combined with multi-agent approach for complex systems," in *2014 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems Proceedings*, Naples, Italy, pp. 1-4, 2014. DOI: 10.1109/EESMS.2014.6923283.

[4] M. Saim, S. Ghapani, W. Ren, K. Munawar, and U. M. Al-Saggaf, "Distributed average tracking in multi-agent coordination: extensions and experiments," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2428-2436, Apr. 2018. DOI: 10.1109/JSYST.2017.2685465.

[5] D. B. Megherbi and V. Malaya, "A hybrid cognitive/reactive intelligent agent autonomous path planning technique in a networked-distributed unstructured environment for reinforcement learning," *The Journal of Supercomputing*, vol. 59, no. 3, pp. 1188-1217, Dec. 2012. DOI: 10.1007/s11227-010-0510-3.

[6] Z. Li, L. Gao, W. Chen, and Y. Xu, "Distributed adaptive cooperative tracking of uncertain nonlinear fractional-order multi-agent systems," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 1, pp. 292-300, Jan. 2020. DOI: 10.1109/JAS.2019.1911858.

[7] D. B. Megherbi and M. Kim, "A hybrid P2P and master-slave cooperative distributed multi-agent reinforcement learning system with asynchronously triggered exploratory trials and clutter-index-based selected sub-goals," in *2016 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Budapest, Hungary, pp. 1-6, 2016. DOI: 10.1109/CIVEMSA.2016.7524249.

[8] H. Lee and S. W. Cha, "Reinforcement learning based on equivalent consumption minimization strategy for optimal control of hybrid electric vehicles," *IEEE Access*, vol. 9, pp. 860-871, 2021. DOI: 10.1109/ACCESS.2020.3047497.

[9] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: a selective overview of theories and algorithms," in *Handbook of Reinforcement Learning and Control. Studies in Systems, Decision and Control*, vol. 325. Springer, Cham, 2021.

[10] J. B. Kim, H. -K. Lim, C. -M. Kim, M. -S. Kim, Y. -G. Hong, and Y. -H. Han, "Imitation reinforcement learning-based remote rotary inverted pendulum control in openflow network," *IEEE Access*, vol. 7, pp. 36682 - 36690, Mar. 2019. DOI: 10.1109/ACCESS.2019.2905621.

[11] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Prentice Hall, 2021.

[12] J. Blumenthal, D. B. Megherbi, and R. Lussier, "Unsupervised machine learning via Hidden Markov Models for accurate clustering of plant stress levels based on imaged chlorophyll fluorescence profiles & their rate of change in time," *Computers and Electronics in Agriculture*, vol. 174, Jul. 2020. DOI: 10.1016/j.compag.2019.105064.

[13] D. Xu and T. Ushio, "On stability of consensus control of discrete-time multi-agent systems by multiple pinning agents," *IEEE Control Systems Letters*, vol. 3, no. 4, pp. 1038-1043, Oct. 2019. DOI: 10.1109/LCSYS.2019.2920207.

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

[15] M. Madera and D. B. Megherbi, "An interconnected dynamical system composed of dynamics-based reinforcement learning agents in a distributed environment: A case study," in *Proceedings of IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, Tianjin, China, pp. 63-68, 2012. DOI: 10.1109/CIMSA.2012.6269597.

[16] J. C. Bol and J. Leiby, "Status motives and agent-to-agent information sharing: how evolutionary psychology shapes agents' Responses to Control System Design," *AAA 2016 Management Accounting Section (MAS) Meeting Paper*, Aug. 2015. DOI: 10.2139/ssrn.2645804.

[17] H. S. Al-Dayaa and D. B. Megherbi, "Reinforcement learning technique using agent state occurrence frequency with analysis of knowledge sharing on the agent's learning process in multi-agent environments," *The Journal of Supercomputing*, vol. 59, no. 1, pp. 526-547, Jun. 2010. DOI: 10.1007/s11227-010-0451-x.

[18] H. S. Al-Dayaa and D. B. Megherbi, "Towards a multiple-lookahead-levels reinforcement-learning technique and its implementation in integrated circuits," *The Journal of Supercomputing*, vol. 62, no. 1, pp. 588-615, Jan. 2012. DOI: 10.1007/s11227-011-0738-6.

[19] Y. Duan, N. Wang, and J. Wu, "Minimizing training time of distributed machine learning by reducing data communication," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 2, pp. 1802-1814, Apr. 2021. DOI: 10.1109/TNSE.2021.3073897.

[20] W. Wang, W. Zhang, C. Yan, and Y. Fang, "Distributed adaptive bipartite time-varying formation control for heterogeneous unknown nonlinear multi-agent systems," *IEEE Access*, vol. 9, pp. 52698-52707, Mar. 2021. DOI: 10.1109/ACCESS.2021.3068966.

[21] D. Bertsekas, "Multiagent reinforcement learning: Rollout and policy iteration," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 2, pp. 249-272, Feb. 2021. DOI: 10.1109/JAS.2021.1003814.

[22] X. Gan, H. Guo, and Z. Li, "A new multi-agent reinforcement learning method based on evolving dynamic correlation matrix," *IEEE Access*, vol. 7, pp. 162127-162138, Oct. 2019. DOI: 10.1109/ACCESS.2019.2946848.

[23] D. B. Megherbi and M. Kim, "A collaborative distributed multi-agent reinforcement learning technique for dynamic agent shortest path planning via selected sub-goals in complex cluttered environments," in *2015 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision*, Orlando: FL, USA, pp. 118-124, 2015. DOI: 10.1109/COGSIMA.2015.7108185.

[24] Megherbi D. B., Malaya, "A hybrid cognitive/reactive intelligent agent autonomous path planning technique in a networked-distributed unstructured environment for reinforcement learning", The Journal of Supercomputing, Vol. 59, Issue 3, p 1188-121, 2012,https://doi.org/10.1007/s11227-010-0510-3.

[25] H. Qie, D. Shi, T. Shen, X. Xu, Y. Li, and L. Wang, "Joint optimization of multi-UAV target assignment and path planning based on multi-agent reinforcement learning," *IEEE Access*, vol. 7, pp. 146264-146272, Sep. 2019. DOI: 10.1109/ACCESS.2019.2943253.

[26] L. Canese, G. C. Cardarilli, L. D. Nunzio, R. Fazzolari, D. Giardino, M. Re, and S. Spanò, "Multi-agent reinforcement learning: A review of challenges and applications," *Applied Science*, vol. 11, no. 11, p. 4948, May. 2021. DOI: 10.3390/app11114948.

[27] S. Zheng and H. Liu, "Improved multi-agent deep deterministic policy gradient for path planning-based crowd simulation," *IEEE Access*, vol. 7, pp. 147755-147770, Oct. 2019. DOI: 10.1109/ACCESS.2019.2946659.

[28] B. Brito, M. Everett, J. P. How, and J. Alonso-Mora, "Where to go next: Learning a subgoal recommendation policy for navigation in dynamic environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4616-4623, Jul. 2021. DOI: 10.1109/LRA.2021.3068662.

[29] C. Liu, F. Zhu, Q. Liu, and Y. Fu, "Hierarchical reinforcement learning with automatic sub-goal identification," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 10, pp. 1686-1696, Oct. 2021. DOI: 10.1109/JAS.2021.1004141.

Min-Suk Kim

received his M.S. in Telecommunication and Networks from the University of Pittsburgh, USA, in 2010. He also received a Ph.D. in Electrical and Computer Engineering from the University of Massachusetts Lowell, USA, in 2016. He was a senior engineer at the Electronics and Telecommunications Research Institute (ETRI) from 2016 to 2020. Since 2020, he has been an assistant professor with the Department of Human Intelligence and Robot Engineering at Sangmyung University, Cheonan, Korea. His research involves Reinforcement Learning, Deep Learning, Edge Computing and Centralized Cloud Computing.

Hwankuk Kim

received his PhD in Information Security from Korea University, Korea, in 2017. He received the B.S. and M.S. degrees in Computer Science and Computer Engineering from Korea Aerospace University in 1998 and 2000, respectively. He is currently an assistant professor at the Department of Information Security Engineering at Sangmyung University. He worked as an Associate Research Engineer at ETRI (Electronics and Telecommunications Research Institute) from 2002 to 2006, and a Manager for Cyber Security Research Team at KISA (Korea Internet and Security Agency) from 2007 to 2020. His research interests include 5G / 6G network security, software vulnerability analysis, IoT security, and security data analysis.