Hierarchical Multi-Agent Reinforcement Learning with Explainable Decision Support for Human-Robot Teams

Aaquib Tabrez* †
Matthew B. Luebbers*
Kyler Ruvane*
University of Colorado Boulder
Boulder, Colorado, USA

Ashley H. Rabin
Kevin W. King
William Gerichs
DCS Corporation
Alexandria, Virginia, USA

Bradley Hayes
University of Colorado Boulder
Boulder, Colorado, USA

ABSTRACT

Effective communication is critical in human-robot teaming, significantly impacting coordination and team fluency, especially in uncertain environments. As part of this communication, providing humans with a clear rationale for an autonomous system's behavior and suggestions can improve collaboration and safety in real-world deployments. This work introduces a novel hierarchical multi-agent reinforcement learning approach for coordinating actions within human-robot teams, generating visual recommendations and explanations for use within partially observable settings. Our proposed approach enables robots to visually communicate their reasoning to human teammates, facilitating the synchronization of environmental uncertainty and improving the interpretability of robot-supplied recommendations. We apply this framework to a dyadic human teaming scenario with robot-provided guidance, leveraging dynamic guidance to nudge participants towards localized roles of leading or following, balancing measures of Granger leadership to improve overall team performance and reduce workload. We propose a set of algorithmic and user study evaluations to assess the impact of the guidance generated by our approach on enhancing team dynamics, fluency, and transparency within multi-robot, multi-human collaborative scenarios.

KEYWORDS

Human-Robot Collaboration, Explainable AI, Multi-agent RL

1 INTRODUCTION AND MOTIVATION

In collaborative tasks within uncertain, dynamic environments, effective communication is crucial for success. This is particularly true in human-robot teams, where bridging the communication gap between human intuition and robotic optimization can enhance teamwork, leveraging the best of each agent and synchronizing notions of uncertainty in partially observable domains [13, 14]. Autonomous agents are well-equipped to navigate probabilistic state spaces, updating their models to select optimal actions in the presence of new information. To facilitate coordinated actions with humans, however, robots must be capable of sharing that evolving knowledge with their human teammates, ensuring both parties remain adaptable to changing conditions, maintaining a synchronized understanding of uncertainties and strategies [15, 19].

Furthermore, previous research in model reconciliation and knowledge sharing within human-robot teams underscores the value of explainability and synchronization of mental models for enhancing trust [18], transparency [4], and overall team efficacy [3, 16]. Tabrez et al. [20] introduced the MARS (Min-entropy Algorithm for Robot-supplied Suggestions) framework, which combined a multi-agent planning algorithm with visual guidance provided to a human teammate, allowing them to smoothly progress towards task completion alongside their robot collaborators. In this work, they found that combining visual insights into environmental uncertainty with explicit robot-provided action suggestions improved trust, transparency, and made human collaborators more independent. In a subsequent work, they also showed that different guidance types given by MARS have a psychological influence on human compliance with action suggestions and how long humans spend scrutinizing that guidance before making a decision [12]. The MARS framework in these works was evaluated in relatively small, discrete domains, and may face computational complexity issues when mapped to large, continuous real-world environments. Furthermore, MARS only supports a single guidance stream, limiting its usefulness to assisting a single human within a multi-agent system.

In this work, we address these limitations, introducing a hierarchical multi-agent reinforcement learning framework with explainable decision support that enables broader operationalization for real-world domains. We additionally aim to investigate the use of tailored visual communication to influence emergent human-human team dynamics within multi-human, multi-robot teams, thus reducing cognitive load and fostering a shared understanding of complex tasks in continuous state spaces.

2 PRELIMINARY: MARS

Multi-Agent RL with Explainable Guidance. We utilize a multi-agent planning algorithm for multi-goal search tasks under uncertainty called MARS (Min-entropy Algorithm for Robot Supplied Suggestions), introduced by Tabrez et. al. [20] as a baseline for our work. In addition to producing robot policies, MARS generates proactive recommendations for human teammates, utilizing two complementary modalities of visual guidance: prescriptive (directly recommending actions) and descriptive (showing state space information to aid in decision-making).

MARS represents uncertainty regarding the location of goals within an environment with a dynamically-updating probability mass function (PMF), a technique commonly used in search tasks [5, 21, 22]. This PMF acts as a shared reward signal for parallel Markov Decision Processes (MDPs): one for autonomous agents (M_R) and another for generating assistive guidance for the human teammate (M_H). MARS solves both MDPs via online reinforcement

*Authors with an asterisk contributed equally to this research as co-first authors.
†Corresponding author: mohd.tabrez@colorado.edu

learning, continually updating the shared PMF in response to agent observations.

Interaction Loop. MARS progresses in a cyclical fashion until all objectives within the environment are achieved. First, M_R is solved to obtain actions for autonomous agents, who act and gather new observations, updating the PMF. The updated PMF is used to solve M_H, and the resultant policy recommendation is communicated to the human. In [20], this is done via an augmented reality-based visualization. The human then acts, recovering a local reward, which updates the PMF again, which is used to again solve M_R, etc. (refer to [20] for a more detailed algorithmic description).

Prescriptive & Descriptive Visual Guidance. MARS communicates its visual guidance using a combination of two modalities: the first is 'prescriptive guidance,' which directly suggests actions to humans (such as holographic arrows for navigation), requiring minimal mental effort but necessitating high trust in the system due to lack of decision-making rationale. The second is 'descriptive guidance' which displays latent environmental information (such as a PMF heatmap), enabling humans to make informed decisions with higher cognitive effort, while offering more flexibility and transparency.

Limitations and Contribution. MARS is a promising framework for integrating human teammates into complex multi-agent robot planners for multi-objective navigation and search tasks, but suffers from scalability issues when confronted with large numbers of agents and high state counts, limiting its applicability for certain real-world robotics domains. In this work, we introduce a spatial hierarchy technique for visual explanation generation, allowing the MARS framework to be tuned to tasks with arbitrary environment size and spatial resolution requirements. We exploit the inherently hierarchical nature of search tasks to transition between levels of state and action abstraction depending on the phase of search, allowing for planning at varying levels of detail (similar to how humans naturally think about search) [1, 7]. This methodology allows the MARS framework to be applied to a much broader class of real-world search scenarios.

Currently, MARS supports single human-in-the-loop interaction, with results showing that guidance type and content have a noticeable effect on the thought patterns of human teammates. Namely, we see that prescriptive guidance induces a passive, automatic, System I [9] style of thinking, limiting adaptability to unexpected changes or flawed recommendations. Conversely, descriptive guidance engages users in active, analytical, System II thinking, encouraging them to plan and make decisions independently. In this work, we extend MARS to manage multiple humans within a team, moving beyond influencing individual human decision-making to influencing emergent human-human team dynamics (such as leader-follower roles) through the provision of tailored, differential guidance. We leverage this novel framework to improve fluency and efficacy of teams involving multiple robots and multiple humans.

3 APPROACH

In this section, we describe our novel hierarchical multi-agent reinforcement learning planner. The novelty of this algorithm is twofold compared to the previous state of the art MARS algorithm in [20]: 1) it introduces a hierarchical structure capable of reasoning over arbitrary environments, which not only makes it scalable to real-world applications, but also enhances the interpretability of guidance [1, 7], and 2) it can provide multiple streams of guidance, expanding its use to multi-human, multi-robot teams.

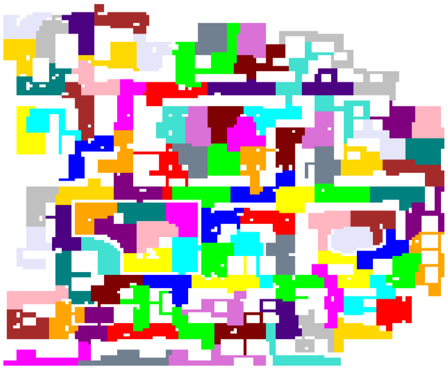

Figure 1: Results of graph partition on 2-dimensional projection of an experimental environment. In this example, the environment is divided into approximately 10,000 grid squares (3m x 3m each), grouped into 100 regions, with impassible obstacles rendered in white.

3.1 Hierarchical MARS Algorithm

At a high level, the hierarchical algorithm functions similarly to MARS as described in [20]. We call this new algorithm H-MARS (Hierarchical Min-entropy Algorithm for Robot-supplied Suggestions). Human and robot Markov Decision Processes (MDPs), encoding the heterogeneous goals and capabilities of each agent class, are solved via online reinforcement learning to generate actions for robot agents and action suggestions for human agents, using a shared, dynamically updating state-wise probability mass function (PMF) to synchronize a notion of likely goal locations between all agents. The algorithm differs, however, in the addition of the ability to group together low-level states into a smaller number of larger regions. H-MARS is capable of dynamically switching between levels of state space abstraction for providing its actions and guidance: considering the entire environment with regions as states, or considering a single region with low-level discretized states (e.g., grid squares). The concept is inherently recursive, and can be extended beyond two levels of spatial resolution: for example, an environment could be divided into regions, which are themselves divided into sub-regions, which are divided into individual states.

To obtain these regions, we discretize our environment into a grid of a desired spatial resolution, and form a graph with grid squares as nodes and edges connecting adjacent, traversable nodes. We then run the METIS graph partition algorithm [10] over this graph, producing contiguous regions of reachable states. To optimize for computational efficiency when running the algorithm in real-time, the number of regions produced should roughly equal the nth root of the total number of states in the environment (for a desired n-level hierarchy). By considering an equal number of states in each phase, the complexity of the combined computation is minimized, reaching a state of Pareto-optimality [2]. An example of this can be seen in Fig. 1, where an environment of 10,000 discrete states is programmatically divided into 100 regions. Assuming a two-level hierarchy, the algorithm progresses through three phases for each time-step, corresponding to swapping the state space, action space and reward function input to MARS between levels of abstraction:

Phase 1 (Local Window Search): The algorithm first considers individual states within a limited distance of each agent. This is to avoid edge cases that would arise by starting Phase 2, involving potentially high-reward actions taking agents to physically nearby states that happen lie across a region boundary. By considering these actions first, we avoid the situation where they receive an outsized reward penalty, normally given to represent the time taken to travel to a separate region.

Phase 2 (Inter-Regional Search): If the tuned reward threshold within Phase 1 is not passed, the algorithm moves on to considering entire regions as single states, with the state-wise PMF used to calculate the expected number of targets to be found per region. The algorithm decides whether it is preferable to stay and search within the current region, or travel to a new, more target-rich region, considering the added movement penalty for taking the time to travel to a separate region, proportional to that region's distance from the current region. If the algorithm decides an agent should move regions, it commands actions or provides action recommendations that path the agent to the nearest edge of the new target region. If the algorithm decides to stay within the current region, it progresses to Phase 3.

Phase 3 (Intra-Regional Search): The hierarchical MARS algorithm now moves to consider the states within an individual region for calculating optimal agent actions, utilizing the PMF value of states in reward calculations, identical to the state space, action space, and reward function of MARS as described in [20]. The phases are repeated every time the global PMF updates in response to the accumulation of agent observations.

4 PROPOSED EVALUATION

We plan to validate the utility and applicability of our Hierarchical MARS framework through a series of algorithmic and user study evaluations. The algorithmic evaluation will measure the empirical scalability of our planner compared to MARS, as well as a state of the art traditional planner. Next, we will evaluate the quality of guidance generated by our planner in an expert feedback user study. Lastly, we plan to assess the impact of our guidance on team performance and emergent leader-follower dynamics in a human dyadic search task conducted in a realistic 3D simulation environment. The environment, implemented in Unreal Engine, offers a large state space and level of realism more comparable to those found in real-world search tasks.

4.1 Algorithm Evaluation

We plan to evaluate the H-MARS algorithm's ability to handle large state spaces by conducting simulation episodes of a multi-objective collaborative search task, with simulated human agents following the system's guidance. Our evaluation will include a comparison of 1) Hierarchical MARS (H-MARS), 2) MARS, as described in [20], and 3) Limited Horizon Multi-objective A* [12], in environments of varying sizes. We hypothesize that H-MARS will demonstrate significantly faster computation times than MARS, particularly as the number of states in the environment increases. Additionally, we hypothesize that H-MARS will outpace Multi-objective A* in terms of task performance metrics, such as the number of targets identified within a specific time frame.

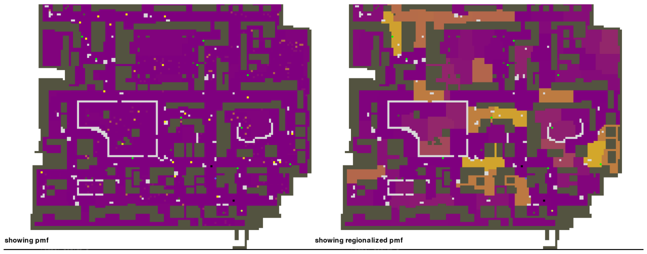

Figure 2: Left: H-MARS descriptive guidance (PMF), coloring each state according to its probability of containing a goal (dark purple to bright yellow). Right: H-MARS descriptive guidance, applied to regions rather than individual states. Heatmap coloring corresponds to the expected number of goals within each region.

4.2 Guidance Evaluation

We propose modifying the visual guidance provided by MARS to leverage the hierarchical phases of the H-MARS more effectively. Our primary insight is that visualizing the PMF in extensive environments with numerous states can overwhelm users (Fig. 2 Left). Additionally, previous research in human navigation shows that users tend to simplify and scaffold complex environments hierarchically [7, 17]. To address both of these criteria, we propose a dual visualization strategy. During Phase 2 (inter-regional search) of the algorithm, we will display a heatmap over regions rather than individual states, using colors to indicate the expected number of goals within each region (see Fig. 2 Right). Prescriptive arrows will then guide users to the boundary of the targeted region. For Phases 1 (local window search) and 3 (intra-regional search), the visualization will focus solely on the PMF of states near the user's current location, presented as a heatmap with prescriptive arrows directing towards a specific target state. By limiting the number of states users need to focus on, regardless of the phase, we expect the guidance to significantly enhance real-time decision-making.

To evaluate our approach to guidance, we plan on conducting an expert-feedback case study (similar to [12]), where participants watch video of a collaborative search task with one of four descriptive guidance types: 1) individual state PMF, entire environment, 2) regional PMF, entire environment, 3) individual state PMF, limited horizon, 4) our dynamic PMF, switching between types 2 and 3 as the MARS algorithm calls for each. We hypothesize the dynamic PMF will rate significantly better than other guidance types on subjective measures of workload, interpretability, and usefulness for decision-making.

4.3 Leader-Follower Evaluation

4.3.1 3D Simulation Environment. We will implement our improved MARS framework in a 3D collaborative mine-defusing simulation environment, implemented in Unreal Engine (Fig. 3). The environment is far larger than the environment used in the Minesweeper game from [20], both in terms of state count (~40,000 compared with 45) and physical area (~360,000 square meters (or ~89 acres)

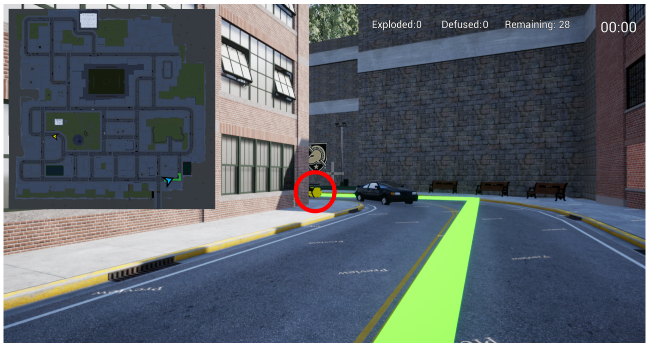

Figure 3: Screenshot from the experimental 3D simulation environment. Here, prescriptive guidance from H-MARS is provided in the form of a green path projected onto the minimap and the environment itself, leading the participant to an undefused mine (circled in red).

compared with ~100 square meters). Unlike the grid-world style 'up, down, left, right' action space from the Minesweeper game, this environment allows participants to move freely throughout a cluttered urban environment, searching visually for mines, with drone teammates providing assistance to speed up the process.

Prescriptive guidance (arrows) will be overlaid onto the 3D gameplay environment, mimicking the augmented reality visualizations developed in [20], as well as shown on a top-down minimap in the corner of the screen, emanating from the player's location (see Fig. 3). Descriptive guidance will be shown on the minimap as well, in heatmap form similar to what is visualized in Fig. 2.

The experimental task within the simulation environment will require human dyads, in order to measure human-to-human team dynamics. Instead of a single player defusing a mine, as in the Minesweeper domain from [20], the game will require both human teammates to enter a radius around a mine at the same time and jointly take defuse actions to progress. These dynamics are drawn from the experimental task in King et al. [11].

4.3.2 User Study. Using this 3D simulation environment, we plan on conducting a human-subjects study to demonstrate the ability of differential guidance to affect human team dynamics. For this study, we will recruit individual participants to play multiple rounds of the dyadic mine defusing game with a research confederate teammate (playing from a separate room to minimize external sources of interference that could affect team functioning, such as conversation or eye contact). The confederate will be given simple instructions to keep gameplay regular between trials (i.e., always to follow the system guidance except to diverge when the location of a mine is confirmed visually).

From the experimental analysis of [11], King et al. find that by labeling the members of otherwise unstructured human dyads as 'leader' and 'follower' in real time (as determined by applying the concept of Granger causality to measure which member's trajectory tends to follow the other [6], producing a measure called Granger leadership [11]), the teammate currently leading can be measured through physiological signaling to be significantly more engaged in the task and expending more mental effort than the follower teammate. Relatedly, the experiment finds that equalizing leadership time between teammates enhances team performance, as it keeps both members actively engaged, while alternating roles reduces mental fatigue by giving each teammate a chance to rest when occupying a less active follower role.

The primary goal of our experiment is to demonstrate that by switching the guidance type provided by an autonomous decision-support system, human teammates in a dyad can be nudged towards 'leader' or 'follower' roles, and that by judiciously switching guidance types per teammate throughout an experimental round, team performance can be enhanced and overall workload reduced. We plan on running three conditions in a within-subjects design, with randomized and counterbalanced ordering of conditions and randomized mine locations between rounds.

Condition 1: Human participant has access to both descriptive (heatmap) and prescriptive (arrow) guidance; confederate has access to prescriptive (arrow) guidance.

Condition 2: Human participant has access to prescriptive (arrow) guidance; confederate has access to both descriptive (heatmap) and prescriptive (arrow) guidance.

Condition 3: Both the participant and the confederate have access to dynamic guidance, with the provision of descriptive guidance (heatmap) switching between teammates via an algorithm that attempts to balance the time spent with each guidance type while conducting guidance switches at minimally distracting times.

We hypothesize that the time periods when the human participant has access to descriptive (heatmap) guidance in Conditions 1 and 3 will be associated with a significantly higher likelihood of measuring the participant as leading via the Granger leadership metric compared with the times when the human participant only has prescriptive (arrow) guidance in Conditions 2 and 3, showcasing the effect guidance type has on team leader-follower dynamics. We also hypothesize that Condition 3 will outperform Conditions 1 and 2 on task performance (the number of mines defused within the time window) and subjective measures of team fluency and contribution to team success by all teammates [8], and hypothesize that participants will report Condition 1 as having the highest workload, followed by Conditions 2 and 3.

5 CONCLUSION

In this work, we present a novel hierarchical multi-agent reinforcement learning approach for partially-observable multi-agent collaborative tasks, called H-MARS. This approach builds on and improves the utility of previous similar approaches [20] by adopting a spatially hierarchical RL structure, making it applicable to large, continuous state spaces. H-MARS informs the generation of visual guidance and explanations, providing human teammates with insight into environmental uncertainty and allowing them to leverage guidance to make informed decisions. We propose a series of algorithmic and user study evaluations to validate our framework, including a dyadic human search task, where robot guidance provided by H-MARS will be used to nudge human teammates towards 'leader' or 'follower' roles. The nudging is achieved by strategically switching guidance types per teammate throughout the task, aiming to balance measures of Granger leadership [11] and thus improve team performance and reduce overall workload.

ACKNOWLEDGEMENT

This work was funded by the Army Research Lab STRONG Program (#W911NF-20-2-0083).

REFERENCES

[1] Salvador Aguinaga, Aditya Nambiar, Zuozhu Liu, and Tim Weninger. 2015. Concept hierarchies and human navigation. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 38–45.

[2] Yair Censor. 1977. Pareto optimality in multiobjective problems. *Applied Mathematics and Optimization* 4, 1 (1977), 41–59.

[3] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. 2021. The emerging landscape of explainable automated planning & decision making. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

[4] Nicholas Conlon, Daniel Szafir, and Nisar Ahmed. 2022. Investigating the Effects of Robot Proficiency Self-Assessment on Trust and Performance. *arXiv preprint arXiv:2203.10407* (2022).

[5] John R Frost. 1997. *The theory of search: a simplified explanation.* Soza Limited.

[6] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* (1969), 424–438.

[7] András Gulyás, József Bíró, Gábor Rétvári, Márton Novák, Attila Kőrösi, Mariann Slíz, and Zalán Heszberger. 2020. The role of detours in individual human navigation patterns of complex networks. *Scientific Reports* 10, 1 (2020), 1098.

[8] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[9] Daniel Kahneman. 2011. *Thinking, fast and slow.* macmillan.

[10] George Karypis and Vipin Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing* 20, 1 (1998), 359–392.

[11] Kevin W King, Stephen M Gordon, and Ashley Rabin. 2023. Granger Leadership in a Novel Dyadic Search Paradigm. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 4889–4894.

[12] Matthew B Luebbers, Aaquib Tabrez, Kyler Ruvane, and Bradley Hayes. 2023. Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming. In *Proceedings of Robotics: Science and Systems*. Daegu, Republic of Korea. https://doi.org/10.15607/RSS.2023.XIX.002

[13] Lanssie Mingyue Ma, Terrence Fong, Mark J Micire, Yun Kyung Kim, and Karen Feigh. 2018. Human-robot teaming: Concepts and components for design. In *Field and Service Robotics: Results of the 11th International Conference*. Springer, 649–663.

[14] Manisha Natarajan, Esmaeil Seraj, Batuhan Altundas, Rohan Paleja, Sean Ye, Letian Chen, Reed Jensen, Kimberlee Chestnut Chang, and Matthew Gombolay. 2023. Human-robot teaming: grand challenges. *Current Robotics Reports* 4, 3 (2023), 81–100.

[15] Stefanos Nikolaidis and Julie Shah. 2012. Human-robot teaming using shared mental models. *ACM/IEEE HRI* (2012).

[16] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The utility of explainable ai in ad hoc human-machine teaming. *Advances in neural information processing systems* 34 (2021), 610–623.

[17] Nestor Schmajuk and Horatiu Voicu. 2006. Exploration and navigation using hierarchical cognitive maps. *Animal Spatial Cognition: Comparative, Neural, and Computational Approaches, MF Brown and RG Cook. p. Available: http://www. pigeon. psy. tufts. edu/asc/Schmajuk/Default. htm* (2006).

[18] Aaquib Tabrez, Shivendra Agrawal, and Bradley Hayes. 2019. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 249–257.

[19] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. 2020. A Survey of Mental Modeling Techniques in Human–Robot Teaming. *Current Robotics Reports* (2020), 1–9.

[20] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. 2022. Descriptive and prescriptive visual guidance to improve shared situational awareness in human-robot teaming. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 1256–1264.

[21] Michał Wysokiński, Robert Marcjan, and Jacek Dajda. 2014. Decision support software for search & rescue operations. *Procedia Computer Science* 35 (2014), 776–785.

[22] Lu Yadong and Zhou Ya. 2015. Optimal Search and Rescue Model: Updating Probability Density Map of Debris Location by Bayesian Method. *International Journal of Statistical Distributions and Applications* 1, 1 (2015), 12.