Adaptive Arbitration for Minimal Intervention Shared Control via Deep Reinforcement Learning

1st Shiyi You
The Department of Automation
University of Science and Technology of China
Hefei, China
ysy3765@mail.ustc.edu.cn

2nd Yu Kang
The Department of Automation
University of Science and Technology of China
Hefei, China
kangduyu@ustc.edu.cn

3rd Yun-Bo Zhao
The Department of Automation
University of Science and Technology of China
Hefei, China
ybzhao@ustc.edu.cn

4th Qianqian Zhang
The Department of Automation
University of Science and Technology of China
Hefei, China
ZQQ789@mail.ustc.edu.cn

***Abstract*—In shared control, humans and intelligent robots jointly complete real-time control tasks with their complementary capabilities for improved performance unavailable by neither side on its own, which is attracting more and more attentions in recent years. Arbitration, as an indispensable part of shared control, determines how control authority is allocated between the human and robot, and the definition of that policy has always been one of the fundamental problems. In this paper, we propose an adaptive arbitration method for shared control systems, which minimizes the deviation from the human inputs while ensuring the system performance based on deep reinforcement learning. We provide humans the maximum assistance with the minimal intervention, in order to balance human's need for control authority and need for performance. We apply our method to real-time control tasks, and the results show that our method achieves high task success rate and shorter task completion time with less human workload, while maintaining higher human satisfaction.**

***Index Terms*—Arbitration, Shared Control, Minimal Intervention, Deep Reinforcement Learning**

I. INTRODUCTION

With the development of artificial intelligence, intelligent robots are capable of building their own behavioral strategies, including goal prediction, strategic planning and action execution, exceeding predefined behaviors. However, complete autonomy is still difficult to achieve, and the shared control of humans and intelligent robots is a feasible solution to the complexity and unpredictability of actual tasks, which is attracting more and more attentions in recent years.

In shared control, humans and intelligent robots jointly complete real-time control tasks with their complementary capabilities, for better performance than their individual control [7] [12]. Take drone landing as an example: humans have greater flexibility in changing factors but it is difficult for them to control in multiple dimensions at the same time, robots have advantages in handling repetitive tasks with high precision and long endurance but it is difficult for them to cope with different complex situations. Shared control combines human inputs and robot actions to address this problem.

Many shared control systems for the tasks determined by humans mainly rely on two components: the inference of human intention which is often not directly available to the robots, and the arbitration between robot actions and human inputs [5] [6] [13]. Arbitration determines how control authority is allocated between the human and robot, and the definition of that policy has always been one of the fundamental problems [1]. The linear combination is a common form of arbitration and has been widely used in many shared control systems [3] [6] [7]. The arbitration weights as the core factors are often predefined by humans, which may not remain optimal for the system in the long term [9]. To cope with this challenge, some other methods calculate the parameters based on the confidence of the intention inference, and when the confidence is high, the human often loses control authority. [14]. These methods exploit the maximum performance of the intelligent robot but are deleterious for tasks that require humans to make the final decision, especially tasks in dynamic and uncertain environments. On the other hand, excessive intervention violates human's preference for more control authority, and may lead humans to resist automatic assistance instead of getting help from it, which weakens the system performance and human satisfaction [2] [3].

In this paper, we propose an adaptive arbitration method for shared control systems, which minimizes the deviation from the human inputs while ensuring the system performance based on deep reinforcement learning. We provide humans the maximum assistance with the minimal intervention, in other words, when intelligent robots intervene for better performance, they should modify human inputs as little as possible to increase their acceptance of assistance. Specifically, we use the long short-term memory network to infer human's intention and calculate the confidence, and use deep reinforcement learning algorithm to estimate the control effect of all actions (the discrete action space or sampling of continuous action space). We set an adaptive threshold based on the confidence

of intention inference and select the action that is closest to the human input among the actions whose control effect exceeds the threshold as the optimal action for execution, in order to balance human's need for control authority and need for performance.

Our main contributions are summarized as follows:

- An adaptive threshold for action selection that maintains optimal in the changing environment.
- A formulation of arbitration for shared control that adheres to the minimal intervention principle and improves the system performance.
- A shared control system that does not require the dynamic model of the controlled system, the human's behavior strategy or other information about the human's ability, which may be necessary in other works but hard to get [15]–[17].

The rest of this paper is organized as follows. In Section II we provide related works. In Section III we describe our method in detail. In Section IV we give experimental processes and results. We conclude in Section V.

II. RELATED WORK

A. Shared Control

Artificial intelligence technology has developed vigorously in recent years and has been applied to many traditional control and automation fields. However, we are far from the goal of replacing human labor with automation. The main reason is that the environment and even the system itself is dynamically changing, and it is difficult to design the system once and for all. Many automated control systems still require humans to continuously and closely interact with intelligent robots in terms of supervision, goal setting, emergency response, etc., and this mode is called shared control.

One of the core challenges in shared control is to assign appropriate control authority to humans and robots to maximize the integration of human intelligence and robot intelligence, that is, arbitration. For example, humans and robots jointly control the speed of the end effector of the robotic arm, and arbitration determines the degree of influence of their respective actions on the final executed action. Reference [1] divides arbitration into four types:

- co-activity, humans and intelligent robots complete different subtasks, such as humans controlling the direction of the robotic arm, and robots controlling the speed of the end effector.
- master-slave, humans and robots have their own autonomy, but when they conflict, humans retain the ultimate authority.
- teacher-student, the intelligent robot is used to train humans, which primarily entails robotic rehabilitation, and the system constantly attempts to reduce the amount of robotic assistance.
- collaboration, humans and robots are equal partners, which is also the main research area of this article.

The linear combination between human and robot strategies is a common form of arbitration, and the arbitration parameter α can be defined as related with the confidence of intention inference c. Reference [4] describes this model as a line graph defined by three parameters $(\theta_1, \theta_2, \theta_3)$: when $c < \theta_1$, the human control the system alone; when $\theta_1 < c < \theta_2$, α is proportional to c; and when $c > \theta_2$, $\alpha = \theta_3$. Although these approaches improve the task performance, they are contrary to the human's preference to be in control and lead to excessive unacceptable interventions and human dissatisfaction. Human satisfaction and their acceptance of robot autonomy are crucial in shared control, which prompts our method to follow the principle of minimal intervention [11].

B. Deep Reinforcement Learning

Many shared control systems model actual tasks as Markov Decision Processes (MDP) or Partially Observable Markov Decision Process (POMDP), which often assume a priori knowledge of the environment dynamics and the human's behavior policy. For example, reference [6] and [3] models shared control as a POMDP with uncertainty over the human's goal. Reference [8] uses POMDP to build a unified framework for the human-in-loop control system, so that the system monitors the state of the human and robot and gives feedback when necessary. The assumption of prior knowledges limits the application of these methods in practical tasks, and the deep reinforcement learning (DRL) methods that can learn strategies in the process of interacting with the environment has unique advantages in eliminating these dependencies. Reference [10] models the confidence and consistency of human feedback by extending deep reinforcement learning, thereby using discrete human feedback to enhance the performance of robots. Reference [9] proposes a shared control framework based on model-free reinforcement learning, which takes the system states and human's commands or inferred goals (if available) as inputs and produces the optimal action that best match human commands. Our method optimizes the artificially specified arbitration parameter in [9] to adaptive parameter that match the system states in real time, and obtains better control effects.

III. METHOD

As shown in Fig.1, the main process of our method is that the human gives control command a_h based on the system state s. The long short-term memory network infers the human's goal g based on a series of human inputs and system motion trajectories (III-A). The deep Q network estimates the control effects r of all actions based on the inferred goal g (III-B). Then the arbitration module selects the optimal action a based on the confidence of intention inference c and the control effects of actions r (III-C). The controlled system executes the optimal action a_{opt} and evolves to the next state.

A. Intention Inference

The long short-term memory network can process the sequence data associated before and after, and can solve the

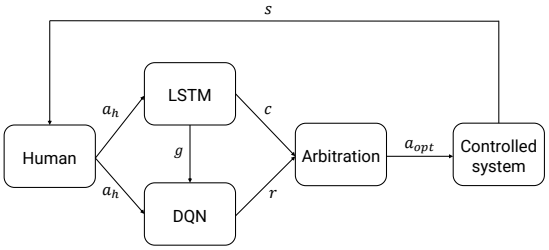


Fig. 1. The block diagram of our method.

problem of gradient disappearance and gradient explosion in general recurrent neural networks, so we use the LSTM network for intention inference. We assume that a set of possible goals G is known (such as all possible landing points in the drone landing task), the human's goal exists in this set but is unknown to the robot. The LSTM network takes a series of human inputs and system motion trajectories as inputs, and predicts the goal g_p. We regard the goal in the known goal set that is closest to the prediction result g_p as the human's goal g, and take the normalized result of the distances between all goals in the goal set and the prediction result as the probability distribution on the goal set. And the confidence of the inference is the maximum probability minus the minimum probability in the probability distribution:

$$c = \max_{g' \in G} p(g'|a_h) - \min_{g' \in G} p(g'|a_h), c \in [0, 1] \tag{1}$$

There are two extreme cases:

- The probability of one goal is 1, and the others are 0. In this case, the inference confidence is 1, that is, the robot is absolutely sure which the human's goal is.
- All goals have equal probability. In this case, the confidence is 0, that is, the robot is completely uncertain which one human's goal is.

B. Control Effects Estimation

The purpose of reinforcement learning is to acquire the optimal strategy so that the sum of the rewards generated by the agent's multi-step actions in the process of achieving the final goal reaches its maximum. The agent selects action a_t at each time step according to the current environmental state s_t and behavior strategy π, that is, $a_t = \pi(s_t)$. The environmental state evolves to s_{t+1} and gives the agent a feedback reward r_t after executing a_t. The transition of this quadruple form will repeat until the system reaches terminal states or the maximum number of transitions, and this process is called an episode. The optimal strategy is to maximize the cumulative reward value $R = \sum_{k=0}^{+\infty} \gamma^k r_{t+k}$ at the end of the episode, where the constant $\gamma \in (0, 1]$ is the discount factor.

Algorithm 1: Minimal Intervention Shared Control via DQN

Initialize experience replay memory $\mathcal{D}$ to capacity N;
Initialize Q-function with random weights θ ;
Initialize target $\hat{Q}$-function with weights $\theta^- = \theta$;
for *episode=1,M* **do**
  for *t=1,T* **do**
    Get environment state s_t and human input a_h ;
    Infer intent and sample action a using Eq.4 ;
    Execute $a_t = a$, observe state s_{t+1} and reward r_t ;
    Store transition (s_t, a_t, r_t, s_{t+1}) in $\mathcal{D}$;
    if s_{t+1} *is terminal* **then**
      for k*=1 to K* **do**
        Sample batch (s_j, a_j, r_j, s_{j+1}) from D ;
        $a'_{j+1} = argmax_{a'} Q(s_{j+1}, a'; \theta)$;
        $y_j = r_j + \gamma \hat{Q}(s_{j+1}, a'_{j+1}; \theta^-)$;
        $\theta \leftarrow \theta - \eta \nabla_\theta \sum_j (y_j - Q(s_j, a_j; \theta))^2$;
      end
    end
    Every C step reset $\hat{Q} = Q$;
  end
end

The optimal policy can be obtained by solving the Bellman equation:

$$Q^{\pi_{opt}}(s, a) = Q^*(s, a) = E_{s'}[r + \gamma max_{a'} Q^*(s', a')|(s, a)] \tag{2}$$

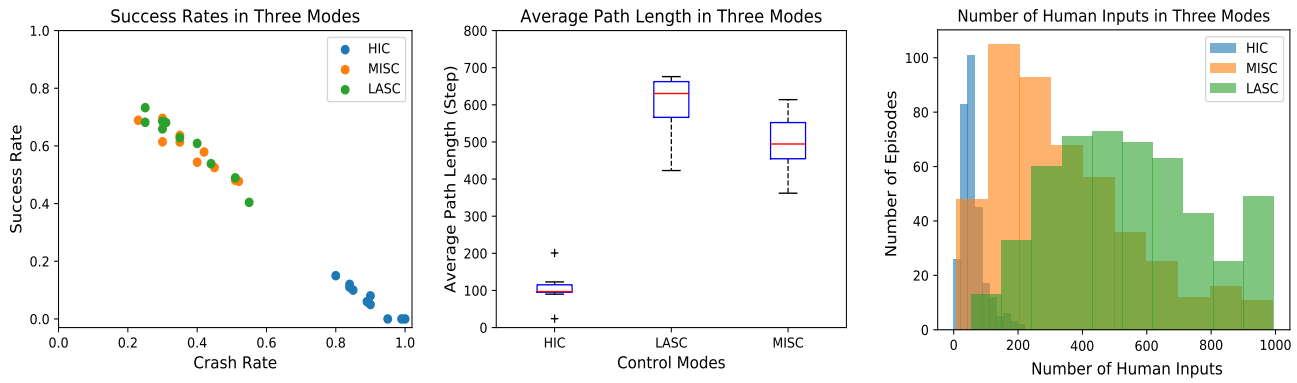
The $Q(s, a)$ is the maximum sum of the discount rewards that can be obtained within the limited steps in the future after performing action a in the state s, which represents the benefit that the action can bring to the current task. The deep Q network (DQN) is a neural network to approximate $Q(s, a)$ [20]. It takes the system states and human commands as inputs, and produces the $Q(s, a)$ for all actions as outputs, using this end-to-end mapping to achieve shared control. Therefore, we use the cumulative reward calculated by DQN as the estimation of the action's control effect.

C. Arbitration

We follow the principle of minimal intervention, that is, when intelligent robots intervene for better performance, they should modify human inputs as little as possible [11]. If the action performed by the controlled system is always far from the human input, the human may no longer trust the system, resulting in a decrease in the information contained in their inputs, which is harmful to intention inference. Therefore, we take the action closest to human input among actions with sufficiently good control effects as the optimal action. The confidence of intention inference determines the adaptive threshold of the control effect. The higher the confidence, the higher the probability that the robot will make the correct decision, so we choose the best action from a smaller range. Formally:

$$A_{threshold} = \{a \in A \mid Q'(s, a) \geq c \times Q'(s, a_{max})\} \tag{3}$$

$$a_{opt} = \underset{a' \in A_{threshold}}{\arg\max} \; f(a', a_h) \tag{4}$$



(a) (b) (c)

Fig. 2. 2(a): The success rates of ten participants performing tasks in three control modes. 2(b): The average steps for each episode of ten participants performing tasks in three control modes. 2(c): The number of human inputs for each episode (the number of keystrokes per episode) of ten participants performing tasks in three control modes.

where A is the discrete action space or sampling of continuous action space and $A_{threshold}$ is the action space calculated according to the threshold to select the optimal behavior a_{opt}. $Q'(s,a) = Q(s,a) - \min_{a' \in A} Q(s,a')$ is designed to prevent the error caused by negative Q values. a_{max} is the action with the highest reward calculated by DQN . The function $f(a', a_h)$ calculates the similarity between action a' and human input a_h. Especially, no input from human will cause the robot to deliver its highest-value action to the controlled system. The overall algorithm is shown in Algorithm 1.

IV. SIMULATION EXPERIMENT

Our method is validated on the simulated Lunar Lander game from OpenAI Gym, as shown in Fig.3. The possible goal set includes three pairs of randomly generated flags on the ground. The human and robot jointly control the three engines distributed on the left, middle, and right sides of the lander to make a smooth landing between the target pair of flags. If the lander crashes into the ground, flies out of the boundary, or fails to land smoothly to the target point within the limited time, the task will fail. The intelligent robot knows the position of the lander and the three pairs of flags, but it needs to infer which pair is the human's goal based on the inputs. The system state vector includes the position, speed, angular velocity of the lander, the angle with the vertical, and whether it touches the ground. The action space is the opening and closing of the three engines. The reward function is to punish speed, tilt angle, distance from the target flag and task failure, and gives the agent a large reward when the task succeeds. The similarity function $f(a, a_h)$ estimates whether the user input a_h and the action a control the same engine or whether they control the lander to move in the same direction, for example, $f((left, on), (right, off)) = 1$, $f((left, on), (left, off)) = -1$.

To estimate the effect of our method, we invited ten participants with an average age of 25 to operate the system in three control modes:

Fig. 3. The Lunar Lander game.

- HIC: human individual control (human-only, no assistance).
- LASC: linear arbitration shared control.
- MISC: minimal intervention shared control.

For LASC, the arbitration weights of robot action and human input are c and $1-c$, respectively. If the fused action is not in the action space, the action with higher weight will be used directly. Each participant operates 40 episodes in advance to familiarize himself with the environment and intelligent robot. In order to facilitate the collection and analysis of data, we assigned the task to the participants to smoothly land the lander in the middle of yellow flags.

The experimental results are shown in Fig.2. Fig.2(a) shows the success rates and the crash rates of ten participants performing tasks in three control modes. The addition of intelligent robots has greatly increased the success rates of the tasks. It is difficult for humans to precisely control the three-dimensional engines to maintain stability when the lander drops, causing the lander to crash into the ground–the crash rates of the ten participants are greater than 0.8. Intelligent robots can effectively control the lander to land slowly, allowing humans to pay more attention to the direction of the lander, thereby greatly increasing the success rate. The success

TABLE I
PARTICIPANTS' RESPONSES (AGREEMENT TO THE STATEMENT) TO SURVEY QUESTIONS.

Survey Questions	LASC	MISC
The assistance from the intelligent robots was helpful.	7.4	7.9
The robot did what I wanted.	8.2	8.2
I accomplished the tasks better with the assistance of robot.	9.2	9.3
I was troubled by the assistance of the robot.	3.6	1.2
I was satisfied with the system.	7.8	8.6

rate of LASC and MISC is almost the same, and the ANOVA result is $F = 0.05, p = 0.8265$, indicating that there is no significant difference between the two. Fig.2(b) shows that the average path length of MISC is shorter than that of LASC, that is, the task is completed faster in MISC mode. The result of ANOVA is $F = 8.65, p = 0.0087$, indicating that the two are significantly different and the gap between them is statistically significant. As shown in Fig.2(c), the number of human inputs per episode (the number of keystrokes per episode) in MISC mode is mostly between 100 and 500, while the number of human inputs in LASC mode is mostly between 300 and 800. We believe that the reason for the difference is that minimal intervention prevents humans from using additional inputs to resist the assistance from the robots, and humans does not need to repeat actions multiple times to make their commands accurately executed, thereby reducing unnecessary workload.

In order to evaluate participants' acceptance of robot autonomy and satisfaction with the shared control system, we asked participants to rate the system performance, as shown in Table I. 10 means strongly agree, 0 means strongly disagree. Participants mostly think that the assistance of the robot is useful and can help them to complete the task better (the first and third rows), and the score of MISC is slightly higher than that of LASC. However, MISC and LASC have exactly the same score on whether the robot did what the participant wanted to do (the second row). We believe that the reason is that the robots and humans have the same ultimate goal but different plans for the specific implementation of each step. When the help provided by the intelligent robot is different from the plan envisaged by the participants, human may use more inputs to fight with the robot autonomy (the fourth row). Humans expect to operate the system with the help of the intelligent robot as a leader and not be disturbed by this kind of help. As shown in the fifth row, MISC gets a significantly higher score for human's satisfaction with the system.

V. CONCLUSION

In this paper, we propose an adaptive arbitration method for shared control systems, which minimizes the deviation from the human inputs while ensuring the system performance via deep reinforcement learning, providing humans the maximum assistance with the minimal intervention. We set an adaptive threshold based on the confidence of intention inference and select the action that is closest to the human input among the actions whose control effect exceeds the threshold as the optimal action for execution, in order to balance human's need for control authority and need for performance. The experiment results show that our method achieves high task success rate and shorter task completion time with less human inputs, while maintaining higher human satisfaction.

ACKNOWLEDGMENT

This work was supported by the National Key Research and Development Program of China (No. 2018AAA0100800, No. 2018YFE0106800), the National Natural Science Foundation of China (61725304, 61673361), the Science and Technology Major Project of Anhui Province (912198698036).

REFERENCES

[1] Losey, Dylan P., et al. "A review of intent detection, arbitration, and communication aspects of shared control for physical human-robot interaction." Applied Mechanics Reviews, vol. 70, no. 1, pp.10804–10804, 2018.

[2] D.-J. Kim, R. Hazlett-Knudsen, et al., "How autonomy impacts performance and satisfaction: results from a study with spinal cord injured subjects using an assistive robot," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 42, no. 1, pp. 2–14, 2011.

[3] S. Javdani, H. Admoni, et al., "Shared autonomy via hindsight optimization for teleoperation and teaming," The International Journal of Robotics Research, vol. 37, no. 7, pp. 717–742, 2018.

[4] Gopinath, D. , S. Jain , and B. D. Argall . "Human-in-the-Loop optimization of shared autonomy in assistive robotics." IEEE Robotics & Automation Letters vol. 2, no. 1, pp.247-254, 2017.

[5] Oh, Yoojin, et al. "Learning arbitration for shared autonomy by hindsight data aggregation." ArXiv Preprint ArXiv:1906.12280, 2019.

[6] Javdani, Shervin, et al. "Shared autonomy via hindsight optimization." Robotics Science and Systems: Online Proceedings, vol. 2015, 2015.

[7] Dragan, Anca D., and Siddhartha S. Srinivasa. "A policy-blending formalism for shared control." The International Journal of Robotics Research, vol. 32, no. 7, pp. 790–805, 2013.

[8] Lam, Chi-Pang, and S.Shankar Sastry. "A POMDP framework for human-in-the-loop system." 53rd IEEE Conference on Decision and Control, pp. 6031–6036, 2014.

[9] Reddy, Siddharth, et al. "Shared autonomy via deep reinforcement learning." Robotics: Science and Systems XIV, vol. 14, 2018.

[10] Z. Lin, B. Harrison, A. Keech, and M. O. Riedl, "Explore, exploit or listen: combining human feedback and policy model to speed up deep reinforcement learning in 3d worlds," arXiv preprint arXiv:1709.03969, 2017.

[11] Broad, Alexander, et al. "Highly parallelized data-driven MPC for minimal intervention shared control." Robotics: Science and Systems XV, vol. 15, 2019.

[12] Abbink, David A., et al. "Haptic shared control: smoothly shifting control authority?" Cognition, Technology & Work, vol. 14, no. 1, pp. 19–28, 2012.

[13] Hauser, Kris. "Recognition, prediction, and planning for assisted teleoperation of freeform tasks." Autonomous Robots, vol. 35, no. 4, pp. 241–254, 2013.

[14] Oh, Yoojin, et al. “Natural gradient shared control.” 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1223–1229, 2020.
[15] Nikolaidis, Stefanos, et al. “Human-robot mutual adaptation in shared autonomy.” Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, vol. 2017, pp. 294–302, 2017.
[16] Fu, Jie, and Ufuk Topcu. “Synthesis of shared autonomy policies with temporal logic specifications.” IEEE Transactions on Automation Science and Engineering, vol. 13, no. 1, pp. 7–17, 2016.
[17] C. Lam, A. Y. Yang, K. Driggs-Campbell, R. Bajcsy and S. S. Sastry, “Improving human-in-the-loop decision making in multi-mode driver assistance systems using hidden mode stochastic hybrid systems.” 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5776-5783, 2015.
[18] Li, Yanan, et al. “Reinforcement learning for human-robot shared control.” Assembly Automation, vol. 40, no. 1, pp. 105–117, 2019.
[19] Tjomsland, Jonas, et al. “Human-robot collaboration via deep reinforcement learning of real-world interactions.” ArXiv Preprint ArXiv:1912.01715, 2019.
[20] Mnih V, Kavukcuoglu K, Silver D. et al. “Human-level control through deep reinforcement learning.” Nature 518, pp. 529–533 , 2015.