Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 180 (2021) 877–886

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry 4.0 and Smart Manufacturing

Human Aspects in Collaborative Order Picking – Letting Robotic Agents Learn About Human Discomfort

Yaxu Niu[a,b], Frederik Schulte[b]*, Rudy R. Negenborn[b]

[a]*Beijing University of Chemical Technology, North Third Ring Road 15, Chaoyang District, Beijing 100029, China*
[b]*Delft University of Technology, Mekelweg 2, Delft 2628 CD, Netherlands*

Abstract

Human aspects in collaboration of humans and robots, as common in warehousing, are considered increasingly important objectives in operations management. This work aims to let robots learn about human discomfort in collaborative order picking of robotic mobile fulfillment systems. To this end, a multi-agent reinforcement (MARL) approach that considers human discomfort next to traditional performance objectives in the reward function of robotic agents is developed. As a first step, we assume a human-oriented assignment problem in which the robotic agents assign orders to human workers at order picking work stations. The results show that among the four evaluated assignment policies, only the proposed MARL policy effectively considers human discomfort. While the approach may need to be refined to obtain near-optimal solutions for the trade-off between humans aspects and efficiency objectives, it also shows a practicable pathway for related problems of human-robot collaboration, inside and outside of warehousing.

Keywords: Order Picking; Robotic Mobile Fulfillment Systems; Human Aspects; Multi-Agent Reinforcement Learning; Human-Robot Collaboration.

1. Introduction

Human-oriented collaboration between humans and robots is widely considered one of the greatest challenges in the final steps of the 4th Industrial Revolution and an anticipated central question of the 5th Industrial Revolution.

*Frederik Schulte. Tel.: +31628305701
E-mail address: f.schulte@tudelft.nl

1877-0509
10.1016/j.procs.2021.01.338

Order-picking in robotic mobile fulfillment systems (RMFS) is one of the applications in which human-robot collaboration is already a pivotal element of today's working reality. Various authors have recognized and addressed the issue with a respective survey paper [2], a conceptual framework for the integration of humans aspects in planning approaches of ordering picking [3], or specific operational models (e.g., [6]) but there remains a lack of decision support approaches that enable robots to learn about human needs and discomfort when working with them. In this work, a multi-agent reinforcement learning (MARL) approach in which robotic agents effectively learn to consider human discomfort next to established objectives such as minimum processing times is proposed. This assumed human-oriented assignment with its human-oriented decisions is illustrated in Fig. 1 based on the layout of the underlying RMFS. For the conducted experimental study, this paper develops four different policies that or commonly deployed in order picking and that only the proposed MARL approach effectively considers the human discomfort during the collaborative order picking process. While it appears probable that the evaluated policy can be further improved in terms of human aspects as well as operational efficiency, the experiments confirm that the MARL approach enables robotic learning with respect to human aspects and (multiple) other objectives. Since the general characteristics of the considered human-oriented assignment problem resemble many other operational problems including human-robot collaboration, the MARL approach also be adopted in different and new problem settings.

Subsequently, this paper presents the related work on the topic (in Section 2), the human-oriented assignment problem in order picking of an RMFS (in Section 3), the proposed MARL approach (in Section 4), the experimental study with results (in Section 5), and a conclusion with open issues for future work (in Section 6).

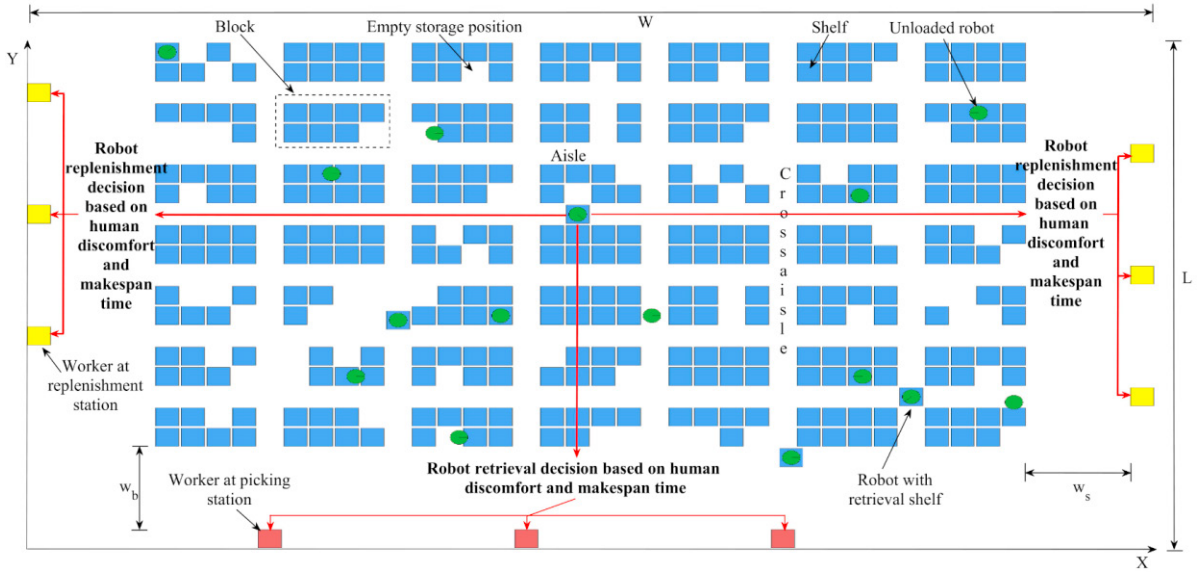

Fig. 1. The layout of the RMFS illustrating robotic assignment decision points in which human discomfort is considered.

2. Related work

The related work for this paper may be distinguished in research on humans aspects on the one-hand side and research concerning decision problems in RMFS on the other side, while there is hardly any work available on human aspects in collaborative order picking of RMFS.

For the integration of human aspects in planning approaches order picking, Grosse et al. [2] provide a systematic literature review, and Grosse et al. [3] propose a conceptual framework for the same purpose. Among specific approaches for human-oriented order picking, Larco et al. [5] suggest a decision support model for suitable storage locations under consideration of empirical models of workers' discomfort.

An overview of research concerning decision problems in RMFS can be found in Table 1. It becomes apparent that most related works focus on different variants of operational assignment problems considering a general assignment

between work stations and robots [15], shelves to storage assignment [11], order and storage assignment [6], velocity-based storage assignment [12], and robots to shelves assignment [8]. Apart from that, workstation location problems [4], pod travel times under different lane characteristics [10], fleet sizing [14], path planning [15], and order batching and shelf sequencing [1] are considered. The most common methods are statistical models, analytical and queuing models as well as optimization models. Only Zhou et al. [14] propose a MARL negotiation scheme and apply it to an order picking example application, and Merschformann et al. [4] develop a discrete event simulation framework for RMFS. Among these references, only Zou et al. [15] consider human aspects using the proxy of the workers' handling speed.

In this paper, we assume a human-oriented assignment problem in RMFS that explicitly takes into account human discomfort, as empirically modeled by [6], in its reward function. Combining a multi-agent model with a (deep) reinforcement learning approach, we specifically model the interaction of robots and humans as well as the robotic learning behavior.

Table 1. An overview of RMFS models with human factors and system decision problems, distinguishing decision problems on strategic (SL), tactic (TL), and operational (OL) levels.

Author	Decision Problem SL	TL	OL	Research objective	Human factors	Method
Zou et al., 2018)			✓	Robot battery charging policies	-	Statistical Model
Zou et al. (2017)			✓	workstation assignment to the robot	Workers' handling speed	Statistical Model
Weidinger et al. (2018)			✓	Shelves storage assignment	-	Statistical Model
Lamballais et al. (2017)	✓	✓		The ratio of length to width and location of the workstation	-	Statistical Model
Wang et al. (2019)	✓			Travel time in multi-deep lanes	-	Statistical Model
M. Merschformann et al. (2019)			✓	Order assignment, shelves selection, and storage assignment	-	Discrete Event Simulation
R. Yuan et al. (2019)			✓	Velocity-Based Storage Assignment	-	Analytical Model
Z. Yuan & Gong (2017)		✓		Optimal number and velocity of dedicated or pooled robot	-	Statistical Model
Zhou et al. (2017)			✓	Robots path planning	-	Multi-Agent Reinforcement Learning Model
Boysen et al. (2017)			✓	Order batching and sequencing of shelves	-	Optimization Model
Roy et al. (2019)			✓	Robots assignment to shelves	-	Statistical Model
This paper			✓	Robot assignment to workstations considering human factors	Workers' discomfort and handling speed	Multi-Agent Reinforcement Learning Model

3. Problem description

In this section, the underlying RMFS is described and the assumptions used in the paper are presented. Then the human-oriented robot assignment problem is defined, both considering system efficiency and human well-being objectives.

3.1. The robotic mobile fulfillment system

The layout of RMFS is shown in Fig. 1, where shelves are organized as rectangular blocks in the storage area storing different items with several bins on each shelf. The workstations are situated along the boundary of the warehouse. Unloaded walking warehouse robots are represented by circles, and the circles with solid-line squares are warehousing robots with a retrieval shelf. Any position in the warehouse can be represented by the coordinates (X, Y). Both aisles and cross-aisle in the storage area are single travel directions to avoid deadlock and congestion.

There are two processes in an RMFS: the retrieval transaction and the storage transaction. These two processes work as following 6 step [16]: (1) The transaction orders arrive and wait in the external queue until it is assigned to the first available robot; (2) The available robot move underneath the shelves from the dwell location toward the

specific retrieval or replenishment shelf; (3) Robots lift the movable shelf and transport it to the designated workstations using the aisles and cross-aisles which arranged along the X-axis and Y-axis respectively; (4) The robot enters the work station buffer and queues for its turn if the worker is busy; (5) The operators fulfill order picking or inventory replenish processes at the workstations; (6) The robot transports the shelf to its previous storage position. Considering that the two processes are similar and retrieval transaction is the most critical process influence the system performance level, only the retrieval process is considered in this paper.

Usually, the ultimate goal is to keep workers and robots as busy as possible to maximize the efficiency of the system. Robots with retrieval shelves prefer to be assigned to a near and fast handling work station. However, due to the differences in workers and previous workload, each worker will have a different discomfort level D_{wi} before our order tasks start. It seems that it is unreasonable to assign a worker who feels discomfort more retrieval tasks, even he/she is efficient. As the pervasive problem in warehouses, reducing discomfort felt by workers not only relates to their well-being but also yields long-term economic benefits through higher productivity, reduced worker health-related costs, absenteeism, and drop-out rate [5]. Human characteristics, such as discomfort and fatigue are often critical factors in warehouse operations, however, are almost always ignored in operational decision making [2]. Therefore, in this paper, the workers' discomfort is considered in the robot assignment problem to determine which work station should be selected for the robot with a retrieval shelf realizing an equal discomfort distribution among the picking workers.

3.2. Human-oriented assignment problem of the robot to workstations

This paper mainly focuses on the assignment of a robot with a retrieval shelf to workstations with different handling-speed, considering both system efficiency and human well-being objectives. For the studied assignment problem, assume that the previous decision problems, such as order assignment and picking shelf selection have been solved. Thus, the next target shelf has already been determined. For the model formulation, some assumptions, which are reasonable in real RMFS, are first listed as follows.

- It is assumed that all the robots are busy and there are always retrieval orders waiting in the external queue.
- The environment of the system is fully observable to each robot, which can be achieved through wireless communication, and wireless communication can also avoid robot collisions with each other.
- Each robot in the system is self-interested and executes its tasks independently.
- Assume that the travel velocity of a robot is constant and ignore the effect of acceleration/deceleration.
- Robots are scheduled based on the First-Come-First-Served (FCFS) policy.
- A shelf has three layers which store different type of products.
- The waiting time in the queue of a work station is constant.
- The fixed orders are single-line orders, which are common and account for a large number in E-commerce orders.
- The picking time of each work station is constant but different from each other.

In a typical RMFS, there always exist numerous unfulfilled order tasks. suppose that R robots are dealing with O fixed sequence orders scattered among different shelves parking in different positions in the storage area, and then will be assigned to pickers with different handling speed T_{w_l} to fulfill the retrieval process. This robot assignment problem is modeled as a Markov Game model, in which all the robots make assignment decisions following a joint policy π, $\pi \in \Pi$ to service a sequence of order tasks.

The robot r_i at the dwell position receives an order task located at the position p_{o_j} and can realize retrieval transaction by assigned to one of the workstations. Let $\mathbf{A} = [0, 1, \cdots, n_W]$ be the set of all the assignment decisions. When n_l is selected in the set $[1, \cdots, n_W]$, it means the robot r_i is assigned to the corresponding workstation w_l, whereas 0 means that the robot is not making allocation decisions at the moment. The assignment decision is based on the system state information that including robot state vector, order information vector, and workstation information vector.

The robot can determine its current position p_{r_i} by scanning the QR code on the warehouse ground. Considering the current time step, the position can be expressed as $p_{r_i}^t$. For each robot, a boolean b is used to indicate task status. If the robot is working on the current order task, $b = 0$, otherwise $b = 1$. Besides, each robot can record the expected

remaining time $\mathbf{RT} = [RT_{r_1}, ..., RT_{r_R}]$ for all robots to complete the current order task, which $RT_{r_i} = T_{r_i}^{o_j} - t_{r_i}^{o_j}$. $T_{r_i}^{o_j}$ is the expected makespan time of current order, which is composed of handling time T_{w_l} of assigned workstation, movement time $T_{d,sh}$ from dwell position to the targeted shelf, and round-trip time $T_{w_l,sh}$ between workstation and shelf storage position. $t_{r_i}^{o_j}$ is the execution time of the current order. And based on the expected makespan time of each order task, individual robot execution time (IRET) can be defined as

$$IRET_{r_i} = \sum\nolimits_{o_j \in \boldsymbol{o}_j} IRET_{r_i}^{o_j} = \sum\nolimits_{o_j \in \boldsymbol{o}_j} T_{d,sh}^{o_j} + T_{w_l,sh}^{o_j} + T_{w_l}^{o_j} \tag{1}$$

Furthermore, the maximum time of all the individual robot execution time of the assigned orders determines the total makespan time (TMT), namely $TMT = \max_{r_i \in R} IRET_i$. From the perspective of system efficiency, an assignment policy is required to minimize the IRET for each robot, without sacrificing other robots' IRET, to minimize the TMT. Based on the above description, the robot state vector is defined as $\mathbf{RS} = [p^t, p_{o_j}, b, \mathbf{RT}]$.

To evaluate the discomfort rating for a product picking at a certain layer on the shelf with specific mass and volume, this paper uses formulation as proposed by Larco et al. [5], expressed in Equation (2).

$$D = d_0 + \sum_{k \in K, k \neq k^*} \alpha^{(\mathrm{k})} L^{(\mathrm{k})} + d_1 HM + d_2 MV + d_3 HV + d_4 MQ + d_5 HQ + \sum_{r \in R, r \neq R^*} d_6 E^{(\mathrm{r})} + IND + \varepsilon \tag{2}$$

where d_0, d_1, d_2, d_3, d_4, d_5 and d_6 are linear coefficients. Low mass, low volume, and small picking quantities are selected as reference values, while products location ($L^{(\mathrm{k})}$, $k = 1, 2, 3$), high mass (HM), medium and high volume (MV, HV), and medium and high pick quantities (MQ, HQ) are regarded as main factors that generate additional workers' discomfort. The dummy variables, which is dented as $E^{(r)}$ in the model, to quantify the individual differences in evaluating discomfort ratings caused by individual traits. The term IND estimates the effect of possible interaction caused by picking at a different level and can also be estimated by a linear model with coefficients $\beta^{(k)}$, $\gamma^{(k)}$, $\lambda^{(k)}$, $\eta^{(k)}$ and $\xi^{(k)}$ to be estimated.

$$\begin{aligned} IND = {} & HM \sum_{k \in K, k \neq k^*} \beta^{(\mathrm{k})} L^{(\mathrm{k})} + MV \sum_{k \in K, k \neq k^*} \gamma^{(\mathrm{k})} L^{(\mathrm{k})} + HV \sum_{k \in K, k \neq k^*} \lambda^{(\mathrm{k})} L^{(\mathrm{k})} + \\ & MQ \sum_{k \in K, k \neq k^*} \eta^{(\mathrm{k})} L^{(\mathrm{k})} + HQ \sum_{k \in K, k \neq k^*} \xi^{(\mathrm{k})} L^{(\mathrm{k})} \end{aligned}$$

Larco et al. [5] also indicated that personal characteristics had little impact on an employees' discomfort level in the studied two real warehouses. And only the interaction between HM and different shelf layers has a significant impact on employee discomfort. This paper adopts the assumption on the order information, shelves, and pickers. And then, the result is directly used to estimate the discomfort rating for a product picking by the following equation

$$D = d_0 + \sum_{k \in K, k \neq k^*} \alpha^{(\mathrm{k})} L^{(\mathrm{k})} + d_1 HM + d_2 MV + d_3 HV + d_4 MQ + d_5 HQ + HM \sum_{k \in K, k \neq k^*} \beta^{(\mathrm{k})} L^{(\mathrm{k})} \tag{3}$$

Therefore, according to Equation (3), the order information vector is composed of the storage layer on the targeted shelf p_{sh}, mass m, and volume v which is defined as $\mathbf{OI} = [p_{sh}, m, v]$. Workstation information vector $\mathbf{WI}$ record the number of completed orders and the cumulate discomfort level of each workstation. From the perspective of discomfort equal distribution among pickers, an assignment policy is required to minimize the cumulated discomfort level for each robot, without increasing other robots' cumulated discomfort. With all required vector, the system state at time t is defined as $\mathbf{S}_t^{r_j} = [\mathbf{RS}, \mathbf{OI}, \mathbf{WI}]$.

Each assignment decision will contribute to the fulfillment time of all order tasks and the discomfort level of the human picker who is assigned to complete the task. It also means that the robot will receive the corresponding penalties on time cost and human discomfort level. The time cost penalty TC_{r_i} is equal to the time each robot has spent since

the last assignment decision. The discomfort penalty is equal to the current discomfort level D_{w_l}, which is the cumulative discomfort of all handled orders. That is, the robot, which is assigned to a picker who feels more discomfort compared to other employees, will have a higher punishment. Moreover, there will be a significant penalty R_p when the idle robot does not make assignment decisions, or when the robot radically makes allocation decisions without completing the current order.

Assuming time cost and discomfort penalties are linear, the penalty function for the robot r_i at sate $\mathbf{S}_t^{r_j}$ with action is given by

$$R_t^{r_j}(\mathbf{S}_t^{r_j}, n_l) = \begin{cases} -R_p, & (b = 1 \text{ and } n_l = 0) \text{ or } (b = 0 \text{ and } n_l \in [1, \ldots, n_W]) \\ -TC_{r_i}^{o_j} - \beta D_{w_l}, & \text{otherwise} \end{cases} \tag{4}$$

β is the weight of the element to evaluate the importance of this item. This paper aims to determine the policy π^* for each robot, that realize equilibrium and maximizes the expected cumulative penalty:

$$J(\pi^*) = \max_{\pi \in \Pi} \{ \sum_{o_j=0}^{O} R_t^{r_i}(\mathbf{S}_t^{r_j}, n_l) \} \tag{5}$$

4. The multi-agent reinforcement learning approach

In this section, a robot-based multiagent reinforcement learning method for the robot assignment problem that aims at obtaining a human-oriented assignment policy considering both system efficiency and the workers' discomfort is presented. In the RMFS, learning in a multi-robot environment is inherently complex. In this paper, based on the independent learning framework, *Q*-learning [9] is used for training each robot and treating other robots as part of the environment during the learning process, which means that each policy is implemented as a separate *Q*-learning process. Specifically, The value of an assignment decision n_l of robot r_j at state $\mathbf{S}_t^{r_i}$ is evaluated by a *Q*-value $Q(\mathbf{S}_t^{r_i}, n_l)$ which represents an estimate of the discounted sum of future penalty. The value of each state-action can be estimated by the iterative update formula defined as

$$Q(\mathbf{S}_t^{r_i}, n_l) \leftarrow Q(\mathbf{S}_t^{r_i}, n_l) + \theta[R_{t+1}^{r_i} + \gamma \max_{n_l \in A} Q(\mathbf{S}_t^{r_i}, n_l) - Q(\mathbf{S}_t^{r_i}, n_l^{'})] \tag{6}$$

where $\gamma \in [0, 1]$ is the discount factor that determines the importance of future rewards; $\theta \in [0, 1]$ is the learning rate that determines what percentage of the old *Q*-value will be fixed by the newly acquired difference between $(R_t^{r_i} + \gamma \max_{n_l \in A} Q_t(\mathbf{S}_t^{r_i}, n_l))$ and $Q(\mathbf{S}_t^{r_i}, n_l^{'})$. In this way, the optimum action for each state is given by

$$\pi^* = \arg\max Q(\mathbf{S}_t^{r_i}, n_l) \tag{7}$$

However, solving Equation (7) has to deal with the "curse of dimensionality" problem caused by enumerating all the combinations of robot location, order information, and workstation information. To solve this computationally hard problem, the neural network is used to approximate the value function, which is expressed as $Q(\mathbf{S}_t^{r_i}, n_l, \theta)$ [7]. θ is the parameters of network parameter trained by stochastic gradient descent. The training batches of M data are randomly sampled from an experience replay buffer which storages the robot's experienced transaction $< \mathbf{S}_t^{r_i}, n_t, R_{t+1}^{r_i}, \mathbf{S}_t^{'r_i} >$, to reduce the correlations between data. The neural network is trained by minimizing the loss function

$$L(\theta)=\sum_{\omega=1}^{M}[(R_{\omega}^{r_i}+\gamma\max_{n_l'\in A}Q(\mathbf{S}'^{r_i}_{\omega},n'_l,\theta^{-})-Q(\mathbf{S}_{\omega}^{r_i},n_l,\theta))^2] \tag{8}$$

Where θ^{-} are the parameters of a target network that periodically copy the parameters θ of a prediction network to stabilize the training process. The $\mathcal{E}$-greedy policy is used to choose actions at each time step. A random assignment is chosen with probability $\mathcal{E}$ to explore the state-action space. The action with maximum Q-value is selected with probability $1-\varepsilon$ to exploit the past transaction data. To realize a high exploration rate to obtain more knowledge of the environment and gradually decrease it over time, the exploration rate is defined as

$$\varepsilon=\varepsilon_{\min}+(\varepsilon_{\max}-\varepsilon_{\min})(T_{\exp loration}-t)/T_{\exp loration} \tag{9}$$

Where t is the current time step and T is the maximum exploration step. The exploration rate decreases linearly from the maximum exploration rate $\varepsilon_{\max}$ to the minimum exploration rate $\varepsilon_{\min}$. Finally, Algorithm 1 compiles all the steps of the proposed human-oriented assignment policy (HOAP) algorithm.

Algorithm 1 Human-Oriented Assignment Policy

Initialize replay memory D to hold N transitions
Initialize action-value function Q with random weights θ
Initialize state $\mathbf{S}^{r_i}$ and action $\mathbf{A}$
for episode=1, Max **do**
  for robot=1, R **do**
    for $t=1,\ T$ **do**
      Select a random action n_t with probability $\mathcal{E}$
      Otherwise select $n_t=\text{argmax}_{n_l\in A}Q^{*}(\mathbf{S}_t^{r_i},n_l;\theta)$
      Execute action n_t
      Observe reward $R_{t+1}^{r_j}$ and next state $\mathbf{S}'^{r_i}_{t}$
      Store experience $<\mathbf{S}_t^{r_i},n_t,R_{t+1}^{r_j},\mathbf{S}'^{r_i}_{t}>$ in D
      Sample random minibatch of experiences $<\mathbf{S}_k^{r_i},n_l,R_{k+1}^{r_j},\mathbf{S}'^{r_i}_{k}>$ from D
      if $\mathbf{S}'^{r_i}_{k}$ is terminal state then
        $y_j=R_{k+1}^{r_j}$
      else
        $y_j=R_{k+1}^{r_j}+\gamma\max_{n'\in\mathbf{A}}\hat{Q}(\mathbf{S}'^{r_i}_{k},n';\theta)$
      Train the Q-network using $(y_j-Q(\mathbf{S}_k^{r_i},n_l;\theta))^2$ as loss
      $\mathbf{S}_t^{r_i}=\mathbf{S}'^{r_i}_{t}$
    end for
  end for
end for

5. Experimental study and results

In this section, the impact of different robot assignment policies such as random assignment policy (RAP), nearest assignment policy (NAP), optimal assignment policy (OAP), and human-oriented assignment policy (HOAP) on human discomfort are evaluated, the simulation experiments are performed on a small size RMFS with shelves

organized as 5×2 rectangular blocks. In terms of the previous workload, the discomfort level of worker wo_1 and worker wo_2 at workstations w_1 and w_2 is set to 10 and 25, respectively. The handling speed of workstation w_1 and w_2 is set as 10s and 25s respectively because of worker's difference. Further parameters are listed in Table 2.

During the experiment, the agents learn for 6000 episodes with the discount factor is set to 0.99, and the exploration rate decreases from 1 to 0.01 in 20000 time steps. The maximum exploration of each episode is set to 15. And the number of nodes of the two layers in the Q network is 128. Fig. 2 demonstrates the convergence of the proposed method for total time and discomfort.

Table 2. RMFS parameters

Parameters	Value	Parameters	Value
the number of robots R	2	width of aisles and cross-aisles w_a , w_{ca}	1 (meters)
number of workstations N_w	2	width of bottom aisles w_b	2 (meters)
the number of orders O	10	width of side cross-aisles w_s	1 (meters)
number of aisles n_a	1	system width W	19 (meters)
number of cross-aisles n_{ca}	2	system length L	8 (meters)
the number of storage positions of the system N	60	width of a square-shaped shelf w_{sh}	0.9 (meters)

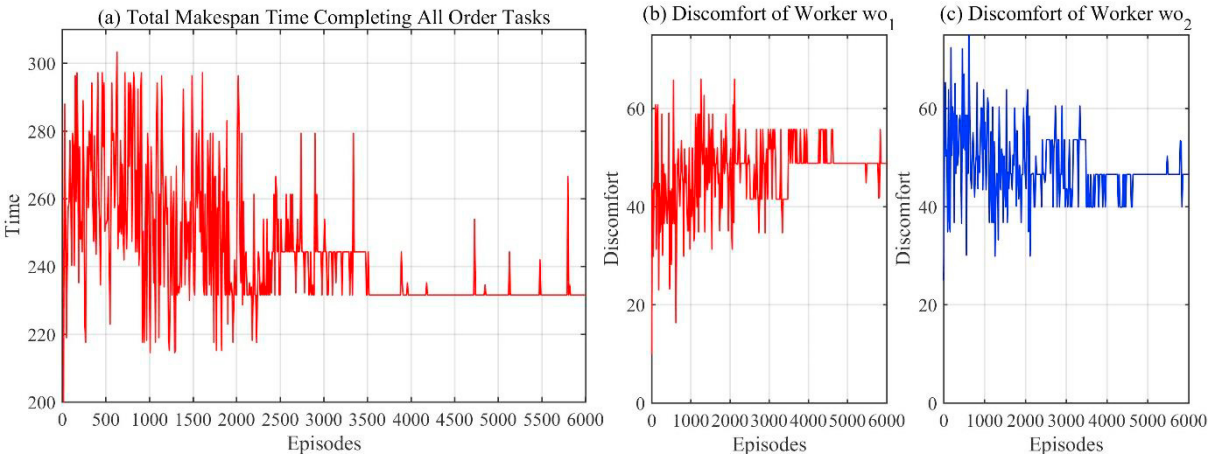

Fig. 2. The convergence of HOAP for (a) total makespan time, (b) discomfort of the worker wo_1 , and (c) discomfort of the worker wo_2

An exhaustive search algorithm is used to find the optimal reference result that only pursues system efficiency to compare the system efficiency of different policies in the robot assignment problem. A DQN-based on the MARL approach, similar to our proposed method but only considers system efficiency, is introduced as a reference method to train the robots to find an efficient robot assignment policy. Moreover, the discomfort difference (DD) between two work stations is measured to evaluate the performance of different assignment policies in terms of discomfort distribution. The experimental results of each assignment policy are shown in Table 3.

In Table 3，it can be seen that, from the perspective of system efficiency, the system can complete all the given orders in 211 seconds by OAP, which is the minimum total makespan time (TMT). Compared to OAP, both NAP and DQN-based MARL are able to complete all the orders in nearly 218 seconds, very close to the minimum TMT. It can be further concluded that, in our experiment, the multi-agent reinforcement learning method can realize a good assignment policy considering system efficiency, which produces the result close to the optimal value. The total makespan time of the HOAP is slightly inferior to the DQN-based MARL and the NAP, but still much better than the average TMT of the RAP, which is 274.06 seconds.

Table 3. The results comparison of assignment policies.

	TMT(s)	D_{w_1}	D_{w_2}	DD
RAP	274.06	40.239	55.032	14.793
NAP	218.27	28.717	66.435	37.718
OAP	211.63	41.718	53.434	11.716
DQN-based MARL	218.93	58.857	36.765	22.092
HOAP	231.54	48.823	46.564	2.259

RAP (random assignment policy); NAP (nearest assignment policy); OAP (optimal assignment policy); HOAP (human-oriented assignment policy)

However, although the total makespan time based on the OAP is the shortest, it results in a significant discomfort difference between the two workstations. Especially, NAP caused up to 37.718 discomfort differences, and DQN-based MARL also led to 22 discomfort differences. This phenomenon can be explained by the axiom that blindly pursuing system efficiency and ignoring employees' discomfort may lead to some employees taking on more order tasks, even in the case when they are more uncomfortable than others. This assignment results in a great difference in the discomfort level among work station workers. Only the HOAP can realize equal discomfort distribution while considering system efficiency. To illustrate the results more clearly, the comparison results of five different assignment policies to total time and discomfort difference of two work stations are shown in Fig. 3.

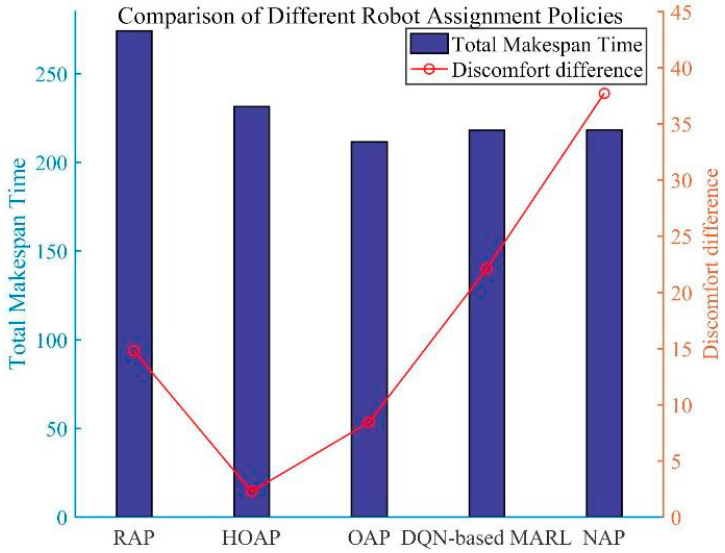

Fig. 3. The comparison of the human-oriented MARL policy (HOAP) and benchmark policies

6. Conclusion

The consideration of human aspects in the collaboration between humans and robots has been widely recognized as one of the major challenges in the era of the 4th Industrial Revolution. Also, human aspects of order picking in warehousing have received significant attention in recent research. However, human well-being has hardly been considered in collaborative order picking (of humans and robots), as common in robotic mobile fulfillment systems. This work has proposed a multi-agent reinforcement learning approach that considers established measures for human discomfort as a penalty in its reward function, for a human-oriented assignment problem in robotic mobile fulfillment systems. The results confirm that the developed approach effectively considers this human discomfort and does so significantly more than four benchmark policies. In this way, our research extends existing work by introducing a new

method for robotic learning in RMFS and by considering human aspects as an objective in order picking of RMFS. Considering that the implemented policy is just a first step, it is particularly interesting to observe the change in the roles of the robots in the collaboration. By learning about human discomfort they could potentially grow into the role of a caring colleague for human co-workers who, for instance, learns and tells humans when they need a break. Moreover, the underlying human-oriented assignment problem resembles a growing amount of related problems, and the proposed MARL approach may therefore also be adopted in those other domains. Nonetheless, it needs to be emphasized that the presented policy is just a initial step, and further refinement will likely lead to better trade-offs of human aspects and traditional efficiency objectives. Future work will, therefore, focus on the development of such advanced policies, also considering different abilities of robots to process human (sensor) signals, and we also aim to explore further use cases in problems of multi human-robot collaboration, inside and outside of warehousing.

Acknowledgments

Funding: This work was supported by the scholarship from the China Scholarship Council (CSC).

References

[1] Boysen, N., Briskorn, D., and Emde, S. (2017). "Parts-to-picker based order processing in a rack-moving mobile robots environment." *European Journal of Operational Research* **262**(**2**): 550–562.

[2] Grosse, E. H., Glock, C. H., and Neumann, W. P.. (2017). "Human factors in order picking: A content analysis of the literature." *International Journal of Production Research* **55**(**5**): 1260–1276.

[3] Grosse, E. H., Glock, C. H., Jaber, M. Y., and Neumann, W. P. (2015). "Incorporating human factors in order picking planning models: Framework and research opportunities". *International Journal of Production Research* **53**(**3**): 695–717.

[4] Lamballais, T., Roy, D., and De Koster, M. B. M. (2017). "Estimating performance in a Robotic Mobile Fulfillment System." *European Journal of Operational Research* **256**(**3**): 976-990.

[5] Larco, J. A., Koster, R. de, Roodbergen, K. J., and Dul, J. (2017). "Managing warehouse efficiency and worker discomfort through enhanced storage assignment decisions." *International Journal of Production Research* **55**(**21**): 6407–6422.

[6] Merschformann, M., Lamballais, T., de Koster, M. B. M., and Suhl, L. (2019). "Decision rules for robotic mobile fulfillment systems." *Operations Research Perspectives* **6**: 100128.

[7] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). "Human-level control through deep reinforcement learning." *Nature* **518**(**7540**): 529–533.

[8] Roy, D., Nigam, S., de Koster, R., Adan, I., and Resing, J. (2019). "Robot-storage zone assignment strategies in mobile fulfillment systems." *Transportation Research Part E: Logistics and Transportation Review* **122**: 119–142.

[9] Sutton, R. S., & Barto, A. G. (1998) "Reinforcement learning: An introduction." *MIT Press*, Cambridge, Massachusetts London, England

[10] Wang, K., Yang, Y., and Li, R. (2019). "Travel time models for the rack-moving mobile robot system." *International Journal of Production Research* **0**(**0**): 1–19.

[11] Weidinger, F., Boysen, N., and Briskorn, D. (2018). "Storage Assignment with Rack-Moving Mobile Robots in KIVA Warehouses." *Transportation Science* **52**(**6**): 1479–1495.

[12] Yuan, R., Graves, S. C., and Cezik, T. (2019). "Velocity-Based Storage Assignment in Semi-Automated Storage Systems". *Production and Operations Management* **28**(**2**): 354–373.

[13] Yuan, Z., and Gong, Y. Y. (2017). "Bot-In-Time Delivery for Robotic Mobile Fulfillment Systems." *IEEE Transactions on Engineering Management* **64**(**1**): 83–93.

[14] Zhou, L., Yang, P., Chen, C., and Gao, Y. (2017). "Multiagent Reinforcement Learning With Sparse Interactions by Negotiation and Knowledge Transfer." *IEEE Transactions on Cybernetics* **47**(**5**): 1238–1250.

[15] Zou, B., Gong, Y., Xu, X., and Yuan, Z. (2017). "Assignment rules in robotic mobile fulfilment systems for online retailers." *International Journal of Production Research* **55**(**20**): 6175–6192.

[16] Zou, B., Xu, X., Gong, Y., and De Koster, R. (2018). "Evaluating battery charging and swapping strategies in a robotic mobile fulfillment system." *European Journal of Operational Research* **267**(**2**): 733–753.