2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

DQN-based on-line Path Planning Method for Automatic Navigation of Miniature Robots

Jialin Jiang[1], *Student Member*, Lidong Yang[2], *Member*, and Li Zhang[1,3,4,5,6], *Fellow*

***Abstract*— Untethered magnetic microrobots with controllable locomotion property and multiple functions have attracted lots of attention in recent years. Owing to the small scale, microrobots with automatic navigation possess a promising perspective for biomedical applications including precise delivery and targeted therapy in confined and narrow space, especially for *in-vivo* scenario. However, the practical working environment for microrobots can be various, dynamic, and complicated, and path planning algorithm applicable for both dynamic obstacle avoidance and planning in maze-like environments still remains a challenge. Furthermore, considering the sizes, different types of microrobots may occupy different proportions of the field of vision. The safe distance between the waypoints and the obstacles needs to be taken into thoughts. In this work, we proposed a reinforcement learning-based strategy capable of real-time path planning for microrobots in different scales. The reference moving direction at each control period is provided by a deep Q network (DQN) according to the local surrounding environment, and the corresponding control magnetic field is generated via a 3-axis Helmholtz coil system. A disturbance observer (DOB) is responsible for the locomotion state observation and direction error compensation. Experiments demonstrate the effectiveness of our proposed strategy using microrobots with different locomotion mechanisms and scales, in both virtual dynamic obstacle environments and channel-like environments.**

I. INTRODUCTION

Remotely actuated microrobots have shown great application potential in biomedicine [1]–[3]. Owing to the small scale, microrobots can noninvasively reach some narrow, confined regions, which are inaccessible to macro conventional medical devices. To date, various actuation strategies have been proposed for wirelessly powering microrobots [4]–[8]. Among these methods, magnetic field possesses the advantages of ideal transparency, biocompatibility, and controllability, and has become one of the favorable power sources in biomedical applications [9]. For instance, researchers have proposed various studies about biosensing [10], targeted delivery [11], and targeted thrombolysis [12] using magnetic microrobots. Furthermore, benefiting from the precise controllability and flexibility of magnetic field, magnetic microrobots are able to perform diverse locomotion mechanisms under different dynamic fields. Magnetic microrobots can be pulled by magnetic field gradient [13], exhibit rotary propulsion [14] and tumbling motion [15] under a rotating magnetic field, or actuated by resonance via an oscillating field [16]. In addition, when multiple microrobot agents form a collective, it can be propelled through the friction asymmetry by introducing a pitch angle to the magnetic field [17]–[19].

To promote the efficiency and accuracy of the navigation of microrobots, autonomy is of great research value. Assisted by advanced control algorithms, researchers have accomplished closed-loop control of microrobots with different structures and moving mechanisms with high precision [20]–[22]. However, in practical working environments, when obstacles exist, direct closed-loop locomotion to the target position is unmet. Thus, an effective obstacle-free path planning method is required.

The existing planning schemes for microrobots can be divided into two categories: off-line planning method and on-line method. off-line planning methods include the iteration-based method (e.g., particle swarm optimization (PSO) [23], [24] and genetic algorithm [25]) and searching-based method (e.g., A-star [15] and rapid-exploring random tree (RRT) [26]). These methods plan the reference trajectories before the navigation based on the images indicating the accessible areas and obstacles distribution of the working environment. The off-line planning methods can provide the optimal, effective obstacle-free trajectory in complicated, multi-branch environments. However, the calculation process could be time-consuming, and the planned path is hard to be modulated in real-time. All these properties mean that the planning results are vulnerable to slight changes of the obstacles, and these methods are not applicable in dynamic environments. on-line planning methods could accomplish real-time moving direction planning according to the target point position and distribution of obstacles. These methods are based on artificial potential field. the microrobots are exposed to the virtual repulsive and attractive forces from the target and obstacle [27], [28]. And eventually, the microrobot would reach the destination while avoiding the obstacles. on-line planning methods are able to deal with dynamic obsta-

This work has received funding support from the Hong Kong Research Grants Council (RGC) (E-CUHK401/20), the RGC/RFS with project Nos. RFS2122-4S03 and RGC/RIF with project Nos. R4015-21. the ITF project MRP/036/18X, the Croucher Foundation grant CAS20403, CUHK internal grants, the Multi-Scale Medical Robotics Center (MRC), InnoHK, at the Hong Kong Science Park, the SIAT-CUHK Joint Laboratory of Robotics and Intelligent Systems. (Corresponding authors: lidong.yang@polyu.edu.hk and lizhang@mae.cuhk.edu.hk)

[1]Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin NT, Hong Kong, China.

[2]Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University (PolyU), Kowloon, Hong Kong, China.

[3]Department of Surgery, The Chinese University of Hong Kong, Shatin NT, Hong Kong, China.

[4]CUHK T Stone Robotics Institute, The Chinese University of Hong Kong, Shatin NT, Hong Kong, China.

[5]Chow Yuk Ho Technology Center for Innovative Medicine, The Chinese University of Hong Kong, Shatin NT, Hong Kong, China.

[6]Multi-Scale Medical Robotics Center, Hong Kong Science Park, Shatin NT, Hong Kong, China.

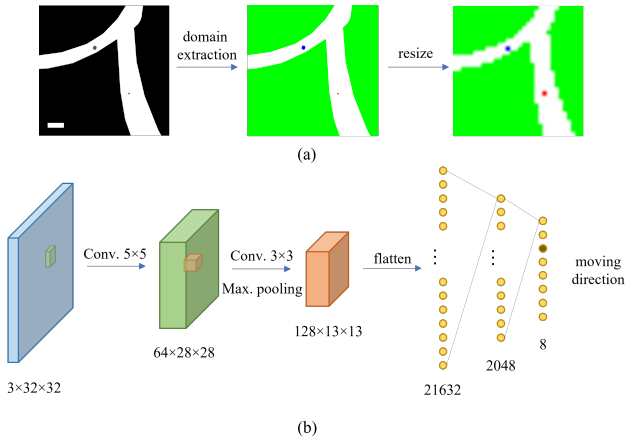


Fig. 1. The processing of the raw image and the network model for DQN. (a) Firstly, the image was captured from a microscopee, the channel, the microrobot, and the target position were extracted and assigned to the different channels of the RGB image. The image was then resized and delivered to the network. The scale bar is 20 microns (b) The neural network contains two convolutional layers and one max pooling layer, the two convolutional kernels are 5×5 and 3×3, respectively. Following the convolutional layers are two fully connected layers. The final output of the network is a 1×8 tensor, representing the Q values of the directions.

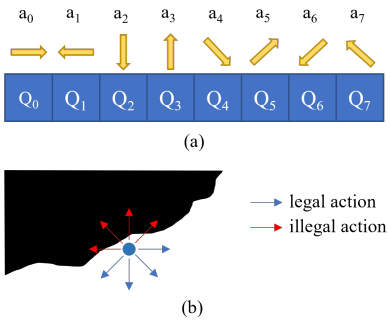


Fig. 2. The action definitions of the agent. (a) The tensor represents the output tensor of the DQN model, $Q_0 \sim Q_7$ are the Q values, and the yellow arrows indicate the specific direction of each action. (b) Illustration of the legal actions and illegal actions. When an action will lead to the collision or collision tendency between the microrobot and obstacle, the action will be illegal and eliminated during the exploration and parameter update.

cles benefiting from their real-time characteristics. However, when the environment is complicated, especially in branchy channel spaces, the virtual resultant forces may not be able the lead the microrobots to the target position.

Learning-based planning method could provide a possible solution to tackle these problems [29]. Mimicking the decision capabilities of human, reinforcement learning (RL) could take images or data matrices as raw inputs, and generate the optimal output with the highest value after training with abundant and effective episodes [30]. Till now, many researchers have explored the possibility of introducing RL algorithms to the navigation applications of microrobots. S. Muinos-Landin *et al*. presented a scheme using Q-learning to navigate a self-thermophoretic microswimmer in a grid space against the disturbance of Brownian motion [31]. Y. Yang *et al*. proposed navigation algorithms based on deep deterministic policy gradient (DDPG) [32] for different types of microrobots. The feasibility of the proposed algorithm was validated with simulation. However, the targeted navigation of microrobots at multiple scales for both dynamic obstacle avoidance and maze-like environment planning still needs exploration. Furthermore, in most studies, microrobots are treated as dots, and only positions are considered. Considering the physical sizes, microrobots with different dimensions may occupy different proportions in the field of vision, the planning should guarantee corresponding safe distances for different scenarios.

In this work, we proposed a DQN-based navigation scheme for full-actuated homogeneous microrobots. The method is applicable for microrobots with different sizes and scales. The input to the deep neural network is the processed images showing the distribution of the surrounding obstacles, the region that microrobots take, and the target position for the local environment. Subsequently, the network would generate the optimal obstacle-free moving direction for the microrobots. The calculation process will take no more than 0.1s, indicating it could be achieved on-line and in dynamic scenarios; furthermore, the planning does not rely on the virtual potential, which means the proposed strategy is applicable in maze-like environments. the feasibility of the navigation scheme is validated via two kinds of microrobots in different scales: magnetic microbead (diameter 5 microns) and microrobot swarms (diameter 500 microns), in both dynamic obstacle situations and real channel-like environments.

II. DQN-BASED NAVIGATION SCHEME

The neural network takes the RGB form images as input, the output is a 1×8 tensor, representing the 8 discrete candidate moving directions. The one with the highest value (Q-value) would be chosen. When the next control period occurs, the input images will be updated with the new position of microrobots and the calculation process will be carried out again, till the microrobot reach the target region.

A. DQN model

The model chosen to fit the Q-value distribution is a convolutional neural network (CNN). The input is a $3 \times 32 \times 32$ tensor converted from the RGB image. The raw image captured from a microscopee was firstly processed to eliminate the noises. After domain extraction, the region of microrobots, obstacles, and target position in the image are assigned to the B channel, G channel, and R channel of the image, respectively. As for microrobots with different scales compared to the field of vision, the region of the microrobots in the tensor will have different sizes, accordingly. Finally, the image was resized to a 32×32 image to fit the input size of the network, as shown in Fig. 1a. After the preprocessing, the image is converted to a tensor and delivered to the network. Then the tensor would be successively convoluted by a $64 \times 5 \times 5$ kernel and a $128 \times 3 \times 3$ kernel. Subsequently, the tensor would be flattened and forwarded to two fully connected layers. The model is shown in Fig. 1b.

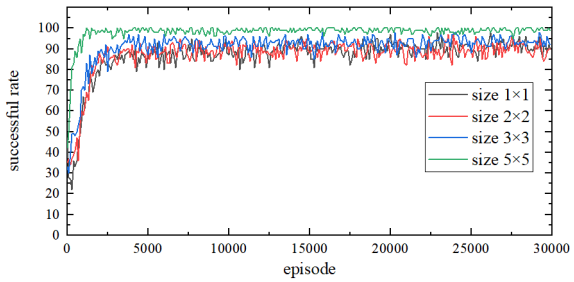


Fig. 3. The searching successful rate of microrobots with sizes of 1×1, 2×2, 3×3, 5×5 pixels are shown in this figure. After the training of 30000 episodes, the success rates are all above 92%.

TABLE I

PARAMETERS OF THE MODEL AND TRAINING PROCEDURE

Parameter	Value
Training episodes N	30000
Batch size B	128
Memory pool size N_m	500000
Target network parameter update frequency f_u	500
Discount factor γ_l	0.9
Learning rate α_l	1e-4
Maximum step for one episode	100
$\varepsilon-$ greedy start possibility	0.1
$\varepsilon-$ greedy end possibility	0.9
$\varepsilon-$ greedy decay factor	200

B. Training of the model

As demonstrated in section II-A, the state of the agent is the tensor containing the distribution of obstacles, the microrobots, and the target point. The action space is a discrete space $A = \{a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$, the specific direction for each action is depicted in Fig. 2a, It is worth noting that to simplify the training process, if one action will cause the microrobot to move too close to the obstacle, the action will be regarded as illegal, and the illegal actions would be eliminated when exploring the environment and updating the policy network.

To train the model to generate the optimal action to reach the target point. The reward for each action is defined as:

$$\begin{cases} -0.1 & if\ action\ =\ a_0 \sim a_3 \\ -0.15 & if\ action\ =\ a_4 \sim a_7\ , \\ 1 & if\ \ |p_r - p_t| < \varepsilon \end{cases} \tag{1}$$

where p_r is the position of microrobot, and p_t is the target point. Since the position update takes a longer distance for actions $a_4 \sim a_7$ than $a_0 \sim a_3$, we give a larger penalty for this situation. When the microrobot is within a pre-defined neighborhood of the target, the navigation is considered finished and a reward is given.

For each episode, the training environment is a small part cut from a large-scale maze-like picture. The start point and target point are randomly generated, and it is guaranteed that at least one effective path exists between the two positions. The model contains two neural networks with the same structure, defined as policy network and target network. The policy network is responsible to choose the action, and target network provides target Q value for parameter update. The Q value gets updated at each step following the equation:

$$Q'_t\left(s_t, a_t|\theta\right) = R\left(a_t|s_t\right) + \gamma \underset{\tilde{a}}{\max} Q_{t+1}\left(s_{t+1}, \tilde{a}|\theta\right) \quad \text{s.t. } a_t, \tilde{a} \text{ are legal} \tag{2}$$

One episode will be terminated when the distance between the microrobot and target reach the threshold or the exploring steps exceed the maximum value. Then the next episode begins and the whole training procedure will start over. The training procedure is summarized in Algorithm 1. The parameters of the model and training are given in Table I.

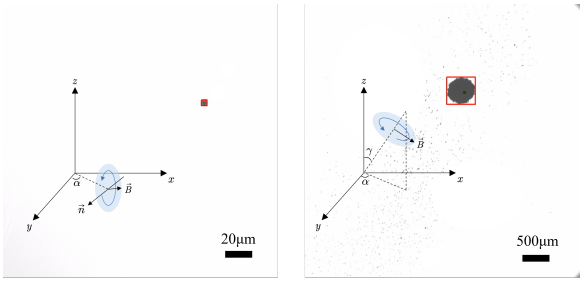
We have trained our DQN model for microrobots with different scales, and the resultant successful rates are shown in Fig. 3. We performed the training for microrobots with the sizes of 1×1, 2×2, 3×3, and 5×5 pixels. After 30000-episode training, the searching successful rate for all the microrobots exceeds 92%.

Algorithm 1 Deep Q learning

Initialize the policy network $Q_{policy}\left(s, a|\theta_p\right)$ and target network $Q_{target}\left(s, a|\theta_t\right)$ with weights θ_p and θ_t. Initialize the memory reply buffer M.
for i in $1 : N$ **do**
  $step$ = 0
  Initialize the start position p_{start}, end position p_{end}.
  while $step < max_step$ **do**
    With a probility slect an action a_t in legal action space that s.t. $a_t = \underset{a}{arg\max} Q_{policy}\left(s_t, a|\theta_p\right)$; otherwise randomly select an action.
    Execute the action with simulation and update the state to s_{t+1}, record the reward r_t.
    Push the transition (s_t, a_t, s_{t+1}, r_t) in the memory buffer.
    if size (memory buffer) >= $batch\_size$ **then**
      Sample random $batch_{size}$ transitions from memory buffer.
      set target value

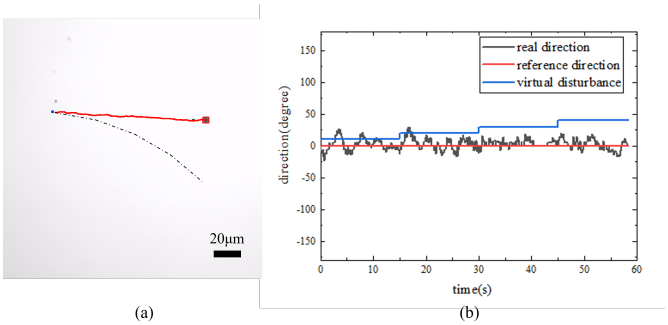
$$y_t = \begin{cases} r\left(s_i\right),\ if\ s_{i+1}\ reaches\ the\ target; \\ r\left(s_i\right) + \gamma \underset{a}{\max} Q_{target}\left(s_i, a|\theta_t\right),\ else \\ \qquad \text{s.t.} a \text{ in leagel space} \end{cases}$$

      Perform gradient decent to update the parameter θ_p for policy network Q_{policy}. on $\left(y_t - Q_{policy}\left(s_t, a_t|\theta_p\right)\right)^2$
    end if
    Every m steps update the parameter θ_t with θ_p.
    $step = step + 1$
    if s_{t+1} reach the target **then**
      Continue
    end if
    $s_i \leftarrow s_{i+1}$
  end while
end for



(a) (b)

Fig. 4. The two types of microrobots with different sizes and scales in the field of vision. (a) SiO_2 microparticle, the diameter is about 5 μm, the actuation magnetic field is a rotating field. The motion direction is related to the yaw angle α. (b) Microrobot swarm consisting of millions of paramagnetic nanoparticles. The motion mechanism of the swarm is based on friction asymmetry due to the pitch angle. The motion speed and direction are governed by the pitch angle γ and yaw angle α.



(a) (b)

Fig. 5. The direction control of the microrobot when virtual disturbance exists. (a) The reference moving direction is 0°, and at 0s, 15s, 30s, 45s, a virtual disturbance of 10°, 20°, 30°, 40° is added to the coils, respectively. The red line is the trajectory of the microrobot, and the black dashed line represents the trajectory of the microrobot without compensation. (b) The real moving direction can be guaranteed to be around the reference value against the disturbance.

III. ACTUATION OF THE MICROROBOTS

In this work, we conduct experiments to validate the proposed algorithm via two types of microrobots. The first one is magnetic SiO_2 microbead. The diameter of the SiO_2 bead is about 5 μm, By applying a rotating magnetic field, the microrobot could perform a tumbling motion. By tuning the yaw angle α and the rotating frequency, the moving direction and speed would be modulated accordingly. The second type is a vortex-like microrobot swarm containing millions of Fe_3O_4 nanoparticles with diameters of $\sim 500\text{nm}$. When exposed to a rotating magnetic field, the paramagnetic nanoparticles will tend to form a vortex-like swarm [33]. By introducing a pitch angle to the rotating surface of the field, the swarm will move owing to the friction asymmetry between the substrate and the swarm. The moving direction and speed are dependent on the yaw angle and pitch angle of the field.

A. Motion characterization of the microrobots

The SiO_2 micro particle and magnetic microswarm are shown in Fig. 4. For the SiO_2 particle, the actuation field is a rotating magnetic field, the rotating surface is perpendicular to the $x-y$ surface, under which the particle could perform a tumbling motion. The yaw angle is defined as the angle between the rotating surface and $y-axis$, which decides the moving direction of the particle. Let p_x and p_y be the position of the microrobot along $x-axis$ and $y-axis$, the dynamics of the microrobot can be defined as:

$$\begin{cases} \dot{p}_x = \mu f \sin(\alpha) + \xi_x \\ \dot{p}_y = \mu f \cos(\alpha) + \xi_y \end{cases}, \tag{3}$$

where μ is a constant to be measured via experiments, f is the rotating frequency of the field, ξ is the unknown dynamics and external disturbances. The subscript x, y represents the component along $x-axis$ and $y-axis$.

As for the microrobot swarm, the locomotion of the swarm is governed by the pitch angle and yaw angle. The dynamics can be expressed as:

$$\begin{cases} \dot{p}_x = h(\gamma)\sin(\varphi) + \varsigma_x \\ \dot{p}_y = h(\gamma)\cos(\varphi) + \varsigma_y \end{cases}, \tag{4}$$

where, ς_x, ς_y are the lumped disturbance and unmodeled dynamics, $h(\cdot)$ is the non-linear function mapping the pitch angle to the moving speed of the swarm. In this work, we only consider the moving direction, thus to simplify the control system, the pitch angle is fixed at 4°, meanwhile the $h(\gamma)$ term would degenerate into a constant. It should be noted that we have abstracted the dynamics of the two types of microrobots to similar forms. The subsequent controller and observer design will be based on equation (3).

Based on the feedback images from the microscope, the position of the microrobot could be extracted after image processing. However, because of the irregularity of the microrobot shape and the uncertainty of the imaging, the position directly segmented from the image may be noisy during the locomotion. Here we utilized a disturbance observer (DOB) to observe the disturbance and the position to filter the chattering.

$$\begin{cases} \dot{\hat{p}}_x = \mu f \sin(\alpha) + L_{x1}(p_x - \hat{p}_x) + \rho_x \\ \dot{\rho}_x = L_{x2}(p_x - \hat{p}_x) \\ \dot{\hat{p}}_y = \mu f \sin(\alpha) + L_{y1}(p_y - \hat{p}_y) + \rho_y \\ \dot{\rho}_y = L_{y2}(p_y - \hat{p}_y) \end{cases}, \tag{5}$$

where ρ_x, ρ_y are the extended states, $L_{x1}, L_{x2}, L_{y1}, L_{y2}$ are the constant factors of the observer (all set to one in experiments), $\hat{p}_x, \hat{p}_y$ are the observed positions along $x-axis$ and $y-axis$. The output of our proposed DQN algorithm is the optimal moving direction, which is the target for the controller. The instant velocity direction $\varphi_i = \text{a}\tan\left(\frac{(\hat{p}_{yi}-\hat{p}_{yi-1})}{(\hat{p}_{xi}-\hat{p}_{xi-1})}\right)$ may amplify the effect of the noise, here we apply mean filter $\bar{\varphi}_i = \frac{1}{m}\sum_{j=0}^{m}\varphi_{i-j}$.

The magnetic field is generated by a 3-axis Helmholtz coil system. The uncertainties of the coils, the installation error of

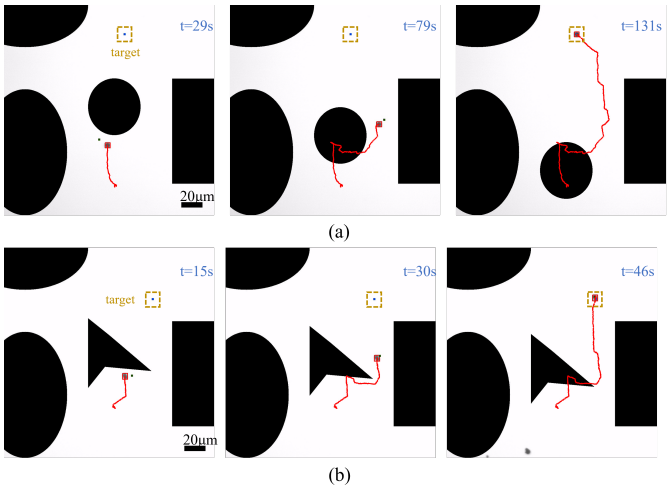


Fig. 6. The avoidance of the dynamic obstacles. In each scenario, three static obstacles and one dynamic obstacle are set in the environment. A round obstacle (a) and an irregular cone-shaped obstacle (b) were assigned to move downwards, and the microrobot managed to avoid the bumping for both situations based on the relative distance and morphology of the surrounding objects.

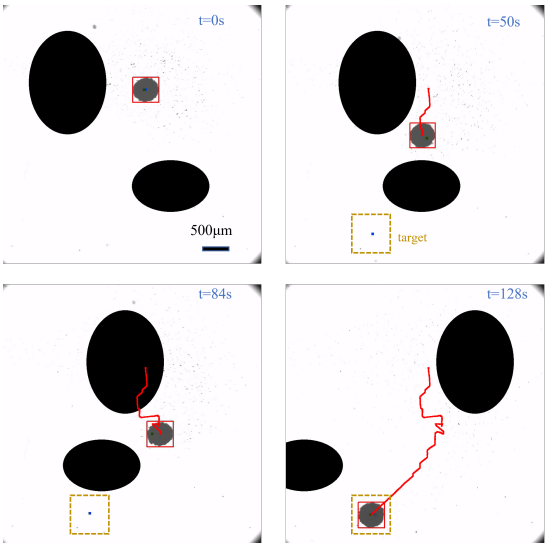


Fig. 7. The navigation of a microswarm with two dynamic obstacles. A paramagnetic microswarm was navigated to the target position marked by the yellow dashed box, and two elliptical obstacles moved toward each other to block the movement. For the original configuration of the obstacles, the swarm tended to cross through between the two obstacles to reach the target. However, with the movement of the obstacles the interval got narrower and not enough for the swarm to pass through, and the algorithm rearranged the path for the swarm to move rightwards and go around behind the obstacle to reach the target.

the imaging system, slight vibration, and the resultant flow could all lead to the deviation of the microrobot moving direction.

$$\varphi_{out} = \varphi + \Gamma, \tag{6}$$

φ_{out} is the yaw angle of the output magnet field from the coils, and Γ is the lumped disturbance. Let φ_{ref} be the reference direction and the feed-forward input of the controller. With the lumped disturbance term to accelerate the convergence and an integration term to eliminate the tracking error, we can obtain the final output angle for each control period:

$$\begin{aligned}\varphi_{out}^{i} =& \varphi_{ref}^{i} + \ k_1\left(\varphi_{out}^{i-1} - \varphi^{i}\right) + \\ & k_2 \Delta t\left(\varphi_{ref}^{i} - \varphi^{i-1}\right) + I_{integrate}^{i-1},\end{aligned} \tag{7}$$

where superscript i and $i-1$ indicate the time instant. k_1 and k_2 are factors, Δt are the control period, $I_{integrate}$ is the integration term.

As depicted in Figure 5, a reference direction of 0° is given, and an virtual disturbance of 10°, 20°, 30°, 40° is added to the coils at 0s, 15s, 30s, and 45s, respectively. The trajectory of the microrobot is marked in Fig. 5a with red line, and the result without the proposed controller is represented by the dashed line. The deviation between the final displacement of the microrobot and the target position is $3\mu m$. Fig. 5b indicates that the real direction of the microrobot fluctuates around the reference direction after the controller enters the stable state against the virtual external disturbance.

IV. EXPERIMENT

To validate the proposed navigation scheme, we conducted experiments using SiO_2 microparticles and paramagnetic microswarms. The navigations were finished in an acrylic tank, and the microrobots were actuated on a silicon substrate, the liquid environment is 0.5% Sodium dodecyl sulfate (SDS) solution. The imaging system is a microscope. A $\times 20$ objective lens was utilized for monitoring SiO_2 microparticle, and a $\times 2$ objective lens was employed for microswarm experiments. A two-computer system was responsible to run the algorithm and generate the desired magnetic field. Computer I processed the feedback images, calculated the parameters of the field, and computer II delivered the parameters to the power amplifiers to actuate the coils. The communication between the two computers was via I/O cards (Model 826, Sensoray, Inc.)

A. Dynamic obstacles avoidance for microrobot

As depicted in Fig. 6, a SiO_2 microbead was navigated to avoid virtual dynamic round-shaped and irregular cone-shaped obstacles. In each scenario, three static virtual obstacles were added. The dynamic obstacles were set to move downwards. Different from the method preplanning the path before the navigation, our proposed method plans the direction based on the surrounding environment real-time. According to the results, the microrobot could effectively avoid the dynamic round obstacle and the cone-shaped structure of an irregular obstacle based on the relative distance and morphology of the surrounding objects.

B. Dynamic obstacles avoidance for microswarm

The dynamic obstacle avoidance experiment was also conducted for microswarms. Compared to the SiO_2 microbead, the swarm occupies a larger proportion of the field of vision, and for the input tensor to the DQN model, the microrobot

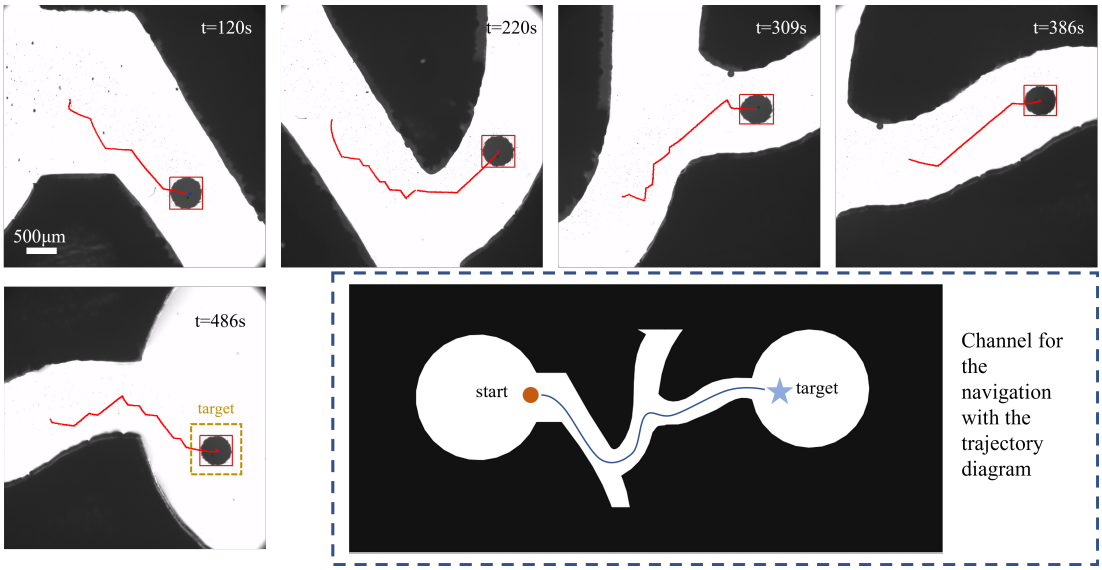


Fig. 8. Targeted navigation of a microswarm in a channel environment. For each field of vision, the local target position was assigned by the operator, and the DQN model would automatically plan the optimal direction for the swarm. According to the result, the swarm could be effectively navigated to the target position in straight channel, curved channel, and cornered channel with a safe distance avoiding bumping into the boundary of the channel. The whole structure of the working channel is depicted at the lower right corner, with the start position, end position, and trajectory diagram marked.

part takes a 3×3-pixel region. The result is depicted in Fig. 7, two elliptical obstacles move towards each other. According to the position of the two obstacles and the trajectory of the swarm, it could be concluded that the original planned path for the swarm was to directly pass through between the two obstacles, since the interval is wide enough, and the path is shorter. However, when t = 50 s, the interval between the two obstacles got narrower and inaccessible for the swarm. Thus, the DQN model started to direct the swarm to move rightwards and went around behind the obstacle to reach the target. The result could indicate that our proposed method is able to rearrange the path according to the local surroundings, and also suitable for dynamic obstacles.

C. Navigation of a microswarm in a channel-like environment

The navigation was also achieved in a channel-like environment. The channel was cut with black nontransparent acrylic materials and glued on a 4cm×7cm silicon substrate. The field of vision of the microscope is 5mm×5mm, which is not enough to monitor the whole channel environment. The movement of the imaging area was finished by an $x-y$ mobile platform. For each local navigation step, the target point is manually assigned. The diameter of the swarm is about $600\mu m$, and the microrobot pixel region in the input tensor of the DQN model is 4×4. The swarm was firstly formed at the left round area, For each navigation subtask, the proposed method could navigate the swarm in straight channel, curved channel, and cornered channel to the target position while guaranteeing a safe distance to the boundary. The whole structure of the channel is exhibited in Fig. 8, the start point, end point, and trajectory diagram for the whole navigation process are also marked.

V. DISCUSSION AND CONCLUSION

In this work, a new DQN-based on-line navigation strategy for homogeneous fully-actuated microrobots with different scales has been presented. In our strategy, A DQN model was employed to generate the optimal moving direction for the microrobot according to the surroundings. The input to the model is a 3-dimensional tensor containing the information of microrobot region, obstacle distribution, and target position. For microrobots with different scales, the according areas in the input tensors would take different sizes. The model was trained in a simulation environment, and the success rate for microrobots with various scales could all exceed 92%. Since the output of the DQN model is the optimal moving direction, and feedforward-based direction control scheme was proposed based on the dynamics of the microrobots. And experimental results indicated that the method was capable to direct the microrobot with high precision against a large disturbance.

To validate the navigation performance of the algorithm in dynamic obstacle environments and channel environments, several experiments for SiO_2 microbead and paramagnetic Fe_3O_4 microswarms were conducted. The avoidance of the dynamic obstacles demonstrated that our proposed method is able to rearrange the trajectory toward the target position dynamically. And also the strategy is applicable in channel-like environments. In the future, we intend to explore the navigation for under-actuated microrobots with heterogeneous structures. Also, we plan to study RL-based navigation when more practical disturbances (e.g., biofluid flow) exist. Moreover, with the aid of medical imaging modalities (*e.g.*, ultrasound imaging), the strategy has the potential to be applied to more clinical scenarios.

REFERENCES

[1] B. J. Nelson, I. K. Kaliakatsos, and J. J. Abbott, "Microrobots for minimally invasive medicine," *Annual review of biomedical engineering*, vol. 12, pp. 55–85, 2010.

[2] M. Sitti, H. Ceylan, W. Hu, J. Giltinan, M. Turan, S. Yim, and E. Diller, "Biomedical applications of untethered mobile milli/microrobots," *Proceedings of the IEEE*, vol. 103, no. 2, pp. 205–224, 2015.

[3] S. Palagi and P. Fischer, "Bioinspired microrobots," *Nature Reviews Materials*, vol. 3, no. 6, pp. 113–124, 2018.

[4] Z. W. Tay, P. Chandrasekharan, B. D. Fellows, I. R. Arrizabalaga, E. Yu, M. Olivo, and S. M. Conolly, "Magnetic particle imaging: An emerging modality with prospects in diagnosis, targeting and therapy of cancer," *Cancers*, vol. 13, no. 21, p. 5285, 2021.

[5] R. Liu, F. Wong, W. Duan, and A. Sen, "Enhanced electrophoretic motion using supercapacitor-based energy storage system," *Adv. Mater.*, vol. 25, no. 48, pp. 6997–7002, 2013.

[6] J. Shi, D. Ahmed, X. Mao, S.-C. S. Lin, A. Lawit, and T. J. Huang, "Acoustic tweezers: patterning cells and microparticles using standing surface acoustic waves (ssaw)," *Lab Chip*, vol. 9, no. 20, pp. 2890–2895, 2009.

[7] B. Qian, D. Montiel, A. Bregulla, F. Cichos, and H. Yang, "Harnessing thermal fluctuations for purposeful activities: the manipulation of single micro-swimmers by adaptive photon nudging," *Chem. Sci.*, vol. 4, no. 4, pp. 1420–1429, 2013.

[8] P. Ryan and E. Diller, "Magnetic actuation for full dexterity microrobotic control using rotating permanent magnets," *IEEE Trans. Robot.*, vol. 33, no. 6, pp. 1398–1409, 2017.

[9] J. J. Abbott, E. Diller, and A. J. Petruska, "Magnetic methods in robotics," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 3, pp. 57–90, 2020.

[10] L. Yang, Y. Zhang, Q. Wang, and L. Zhang, "An automated microrobotic platform for rapid detection of c. diff toxins," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 5, pp. 1517–1527, 2019.

[11] X. Yan, Q. Zhou, M. Vincent, Y. Deng, J. Yu, J. Xu, T. Xu, T. Tang, L. Bian, Y.-X. J. Wang, *et al.*, "Multifunctional biohybrid magnetite microrobots for imaging-guided therapy," *Science robotics*, vol. 2, no. 12, p. eaaq1155, 2017.

[12] Q. Wang, D. Jin, B. Wang, N. Xia, H. Ko, B. Y. M. Ip, T. W. H. Leung, S. C. H. Yu, and L. Zhang, "Reconfigurable magnetic microswarm for accelerating tpa-mediated thrombolysis under ultrasound imaging," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 4, pp. 2267–2277, 2022.

[13] Z. Zhang, F. Long, and C.-H. Menq, "Three-dimensional visual servo control of a magnetically propelled microscopic bead," *IEEE Trans. Robot.*, vol. 29, no. 2, pp. 373–382, 2012.

[14] L. Zhang, J. J. Abbott, L. Dong, B. E. Kratochvil, D. Bell, and B. J. Nelson, "Artificial bacterial flagella: Fabrication and magnetic control," *Applied Physics Letters*, vol. 94, no. 6, p. 064107, 2009.

[15] Z. Yang, L. Yang, and L. Zhang, "Autonomous navigation of magnetic microrobots in a large workspace using mobile-coil system," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 6, pp. 3163–3174, 2021.

[16] H.-W. Tung, M. Maffioli, D. R. Frutiger, K. M. Sivaraman, S. Pané, and B. J. Nelson, "Polymer-based wireless resonant magnetic microrobots," *IEEE Transactions on Robotics*, vol. 30, no. 1, pp. 26–32, 2014.

[17] J. Yu, B. Wang, X. Du, Q. Wang, and L. Zhang, "Ultra-extensible ribbon-like magnetic microswarm," *Nature communications*, vol. 9, no. 1, pp. 1–9, 2018.

[18] J. Yu, T. Xu, Z. Lu, C. I. Vong, and L. Zhang, "On-demand disassembly of paramagnetic nanoparticle chains for microrobotic cargo delivery," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1213–1225, 2017.

[19] L. Yang, J. Yu, S. Yang, B. Wang, B. J. Nelson, and L. Zhang, "A survey on swarm microrobotics," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1531–1551, 2022.

[20] Z. Yang, L. Yang, and L. Zhang, "3-d visual servoing of magnetic miniature swimmers using parallel mobile coils," *IEEE Transactions on Medical Robotics and Bionics*, vol. 2, no. 4, pp. 608–618, 2020.

[21] A. Oulmas, N. Andreff, and S. Régnier, "3d closed-loop swimming at low reynolds numbers," *Int. J. Robot. Res.*, vol. 37, no. 11, pp. 1359–1375, 2018.

[22] X. Du, M. Zhang, J. Yu, L. Yang, P. W. Y. Chiu, and L. Zhang, "Design and real-time optimization for a magnetic actuation system with enhanced flexibility," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 3, pp. 1524–1535, 2021.

[23] L. Yang, Y. Zhang, Q. Wang, K.-F. Chan, and L. Zhang, "Automated control of magnetic spore-based microrobot using fluorescence imaging for targeted delivery with cellular resolution," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 1, pp. 490–501, 2020.

[24] Q. Wang, L. Yang, and L. Zhang, "Micromanipulation using reconfigurable self-assembled magnetic droplets with needle guidance," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 2, pp. 759–771, 2022.

[25] H. Xie, X. Fan, M. Sun, Z. Lin, Q. He, and L. Sun, "Programmable generation and motion control of a snakelike magnetic microrobot swarm," *IEEE/ASME Trans. Mechatron.*, vol. 24, no. 3, pp. 902–912, 2019.

[26] L. Zheng, Y. Jia, D. Dong, W. Lam, D. Li, H. Ji, and D. Sun, "3d navigation control of untethered magnetic microrobot in centimeter-scale workspace based on field-of-view tracking scheme," *IEEE Transactions on Robotics*, 2021.

[27] H. Kim, U. K. Cheang, and M. J. Kim, "Autonomous dynamic obstacle avoidance for bacteria-powered microrobots (bpms) with modified vector field histogram," *PloS one*, vol. 12, no. 10, p. e0185744, 2017.

[28] J. Lee, X. Zhang, C. H. Park, and M. J. Kim, "Real-time teleoperation of magnetic force-driven microrobots with 3d haptic force feedback for micro-navigation and micro-transportation," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1769–1776, 2021.

[29] L. Yang, J. Jiang, X. Gao, Q. Wang, Q. Dou, and L. Zhang, "Autonomous environment-adaptive microrobot swarm navigation enabled by deep learning-based real-time distribution planning," *Nature Machine Intelligence*, vol. 4, no. 5, pp. 480–493, 2022.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[31] S. Muiños-Landin, A. Fischer, V. Holubec, and F. Cichos, "Reinforcement learning with artificial microswimmers," *Sci. Robot.*, vol. 6, no. 52, 2021.

[32] Y. Yang, M. A. Bevan, and B. Li, "Micro/nano motor navigation and localization via deep reinforcement learning," *Adv. Theory Simul.*, vol. 3, no. 6, p. 2000034, 2020.

[33] J. Yu, L. Yang, and L. Zhang, "Pattern generation and motion control of a vortex-like paramagnetic nanoparticle swarm," *Int. J. Robot. Res.*, vol. 37, no. 8, pp. 912–930, 2018.