2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Skill Learning in Robot-Assisted Micro-Manipulation Through Human Demonstrations with Attention Guidance

Yujian An, Jianxin Yang, Jinkai Li, Bingze He, Yao Guo, Guang-Zhong Yang, *Fellow*, *IEEE*

***Abstract*— For the development of robotic systems for micro-manipulation, it is challenging to design appropriate control strategies due to either the lack of sufficient information for feedback or the difficulty in extracting subtle yet critical visual features. With the same system under the teleoperated mode, however, human operators seem to be able to complete the task more successfully with an inherent motion and control strategy. The extraction of implicit human attention during the task and integration of this with robot control could provide crucial guidance in the design of feature extraction and motion control algorithms. In this paper, a micro-assembly task of miniature thin membrane sensors is considered. For human demonstrations, we collected data from repeated tests performed by ten operators following three motion strategies. The human attention during the task is explored according to the coordinates of the eye gaze, and then a neural network with gaze-guided attention is trained to segment the visual Region of Interest (ROI). After quantitative evaluation of operator results in terms of success rate, efficiency, reset time, and the Index of Pupillary Activity (IPA), an optimized motion strategy based on the "palpation" framework was derived. Consequently, we apply this strategy to automated tasks and achieve superior results than human operators, showing an average task completion time of 34.8±5.9s and a success rate of over 90%.**

I. Introduction

Visual perception plays a significant role in macroscopic and microscopic robotic systems [1]–[4]. Accurate extraction of critical features is essential for enabling automatic control of robotic systems. For instance, optimized motion trajectories for completing a specific task can be planned and executed by the robot based on the extracted key information [3], [5]. However, critical information extraction in robot-assisted micro-manipulation remains challenging, especially when the accessible information is insufficient or incomplete for performing such operations [6], [7]. Differently, human beings are good at discriminating the key information from the cluttered background and are able to explore latent perceptual cues from auxiliary movements, i.e., palpation and repeat attempts. Hence, it would be valuable to learn human skills and devise the control strategy of the robots accordingly [8]–[10]. In other words, robots are expected to not only learn the attention mechanism of human beings to focus on critical visual features but also imitate the optimized human behaviors that can compensate for information loss in motion planning.

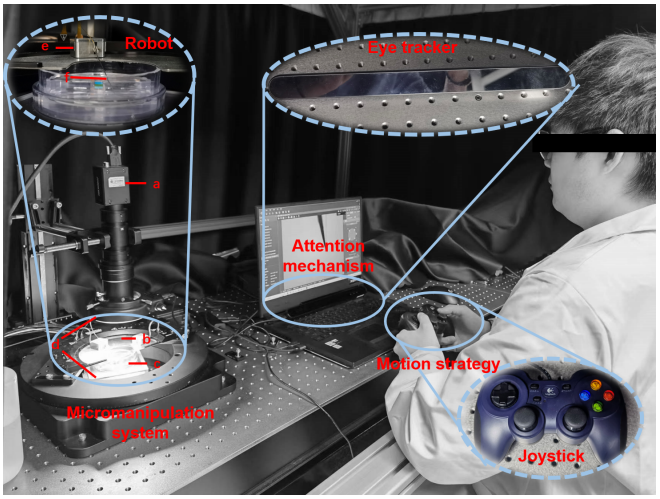

Fig. 1. Experimental scenario for the micro-manipulation task teleoperated by a human operator. The gaze point on the screen is detected by a table-mounted eye tracker, and control inputs are recorded from the joystick. The robotic system includes components a) a top camera for capturing the probe and the sensor, b) a robot platform, c) a pool for deploying the sensor, d) top and bottom light sources, e) micro-robots that control the movement of the probe, and f) a probe.

In this paper, we aim to enable a robot to learn a motion strategy for a 3D micro-manipulation task from operators' behavioral patterns. Based on the image captured from a top-view microscope, a micro-robot carrying a probe needs to pass its tip through a tiny engaging hole of a thin membrane sensor floating on the water surface as encountered in applications such as micro-electrode assembly for neuro-interfaces. Such 3D microscopic tasks can commonly acquire the positional relationship through a multi-camera microscopic system, which usually consists of a top view global camera and a side view local camera [11], [12]. However, the floating micro-sensor is less than 5 μm thick, which is too thin to be distinguished in the side view. Thus, it is challenging to obtain depth information of both the probe and the sensor directly through the top view camera. Since humans could perform the micro-manipulation tasks better under such conditions, we aim to use this as an exemplar to explore the extraction and integration of implicit human behavior for formulating the motion strategy of the robot.

Eye gaze, as an important way for humans to interact with the world, can convey rich information about human attention and mental state [13], [14]. The recording and analysis of

This work was supported by Shanghai Municipal Science and Technology Major Project 2021SHZDZX, and also in part supported by the Science and Technology Commission of Shanghai Municipality under Grant 20DZ2220400.

Y. An, J. Yang, J. Li, B. He, Y. Guo, and G.-Z. Yang are with the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China. ({yujianan, jianxinyang, lijinkai, hebingze, yao.guo, gzyang}@sjtu.edu.cn).

Corresponding authors: *Yao Guo, Guang-Zhong Yang*.

eye movement can provide an essential reference for the robot that performs the same task [15], [16]. Based on these cues, we use in this study a table-mounted eye tracker to capture the operators' eye movements and pupillary response during the tasks. More specifically, their fixation heat maps and trajectories are used to visualize the distribution of their attention. Besides, the real-time pupil diameters of the operators are also recorded, and a metric named the Index of Pupillary Activity (IPA) is calculated to reflect the operators' mental state [17], [18].

Since the sensor is susceptible to noise interference, the engaging hole quivers during assembly, making trajectory planning more difficult. Moreover, the key point for the control strategy designing, i.e., the engaging hole of the sensor, is challenging to extract from the whole image since its diameter is smaller than 25 μm (40 pixels in the image). To this end, we proposed an end-to-end network that can extract the accurate position of the engaging hole from the microscopic image. Since the fixation points while performing the tasks indicate the attention area and the information screening extraction strategy [19], [20], it is a natural supervisory label based on human consciousness [21]–[23]. Therefore, we utilized the fixation points during training, making the network approximate the human attention mechanism, thus boosting the detection accuracy.

In this study, we recruited ten operators to teleoperate with the robotic system to complete the micro-manipulation task, while observing the real-time visual feedback on the monitor. Each operator was asked to repeat the experiment using three different motion strategies for a total of 15 times. For task assessment, evaluation metrics, such as the time spent, success rate, excellence, and reset time, were used. Eye trackers have also been introduced to collect the visual search and attention data to examine whether individual differences bias the results. Finally, we applied the motion strategy with the best comprehensive development of the index evaluation to the automatic mode. We obtained an average assembly time of 32.2$\pm$6.4s and a success rate of over 90%, which is even better than that of human operators. The main contributions of this paper are:

- By exploring the attention mechanisms of human operators, we propose a neural network with gaze guidance to segment the ROI in real-time, which can effectively exclude irrelevant information in low-interest areas.
- Inspired by human operators, an optimized control strategy called "palpation" is proposed for compensating the missing depth information in the micro-assembly of the sensor object.
- Through the skill learning from human demonstrations, the automated system can not only exhibit human-like locomotor patterns but also outperform human operators in terms of task efficiency and success rate.

II. Robot-Assisted Micro-Manipulation

In this section, we first introduce our microscopic vision and robotic systems for micro-manipulation and then describe the teleoperated experiments by human operators for robot skill learning in this paper.

A. Hardware System

The system designed for this micro-manipulation task is shown in Fig. 1, which is optimized based on our previous research [24]. Real-time images are collected under a HIKROBOT MV-CH120-11UC RGB camera (resolution: 4096$\times$3000) with 1x eyepiece and 5x objective lens. This camera set provides sufficient depth of field and Field of View (FoV) for experiments. It ensures that the operation of all objects can be carried out comfortably without worrying about loss of focus or limited range of motion. It is worth mentioning that this camera is placed vertically downward to collect top-view information. The depth information, the relative depth relationship between objects, cannot be directly observed due to tiny size.

The robot used is called Mibot (IMINA Technologies, Switzerland), which has independently driven 4 Degree-of-Freedoms (DoFs), namely X, Y, Z (three directions of axes in the earth coordinate), and R (rotate around the axis of mibot). Mibot has a freely adjustable speed range from 1 to 1000 microns per second and high displacement accuracy: $50nm$ in the X and Y directions and $120nm$ in the Z direction. However, the descent of the Mibot robotic arm is not vertically downward but a circular movement around the rear point. Therefore, the probe's descent will accompany a retreat, requiring forward compensation to restore the horizontal position. This descent-forward cycle dramatically increases the complexity of the operation.

B. Micro-Manipulation Task

The micro-manipulation task in this experiment is an assembly task of passing a probe tip carried by a robot through a hole for subsequent sensor implantation. The operator controls the robot's movement through a joystick and can observe the position of the hole and the probe through the image. The probe tip is $5\mu m$ wide and the target hole is at the head of a sensor up to $5\mu m$ thick and $50\mu m$ wide with a diameter of $25\mu m$. The sensor floats on the gas-liquid interface, which the camera focuses on.

The ten recruited operators include three skilled operators familiar with the system and seven novice operators who need more awareness of the system. Novice operators can spend 10 minutes getting acquainted with the system before officially starting the test. During this time, they can make free trials as needed. For operators without any joystick experience, we extended it to 20 minutes to avoid losing the objectivity of the data due to operator nervousness.

The three motion strategies adopted by the operator are: A) Keep the probe tip directly above the hole at all times, and repeat the descent and forward compensation until task completion. B) Finish the descent in any selected area, then push the probe tip into the hole for fine adjustment. C) The descent is completed at the rear end of the target, followed by "palpation" detection to gauge the exact microscale sensor position: Contact was considered successful when

the electrode could follow the probe tip in synchronized reciprocating movement. Then, push forward into the target hole to complete the assembly.

Each operator was asked to complete the task for five times with each motion strategy in the experiment. We set no time limit for a single teleoperation. One is to relieve the operator's psychological pressure. The other is that some novices may become proficient after a few long-term operations. Sacrificing the first few attempts to "learn from failure" aligns with human psychology, which we do not want to interfere with.

III. Robot Skill Learning

In this section, the learning framework based on eye-tracking techniques is proposed. First, the method of mental state assessment through eye movement analysis is presented. In addition, we propose a new neural network used to obtain regions of interest based on human attention guidance.

A. Eye Movement Analysis

In this work, a 7invensun A3 (7invensun Technology Co., Ltd., Beijing, China) table-mounted eye tracker is applied during the tasks to record the real-time eye movements of the operators. The eye tracker is placed below the display screen, on which the image captured by the camera is displayed. Before the start of each task, the eye tracker is calibrated to the operators' eyes to guarantee the data quality. The supporting software, aSee Studio, is used to process and export the eye-tracking data, which includes the information on original gaze points, fixations, saccades, pupil diameters, etc. Based on these data, we conduct further analysis on the attention and cognitive load of the operators when performing the micro-manipulation tasks.

To analyze the operators' attention, we extract the coordinates and durations of the fixation points and plot the gaze trajectory and heat map using these data, as shown in Fig. 2. For the trajectory, the numbers on the fixation points indicate their chronological order and a larger size refers to a longer fixation duration. For the heat map, a red/blue region respectively indicates a higher/lower density of the fixation points. It should be noted that the fixation duration for each task has been normalized before plotting the figures, which ensures an accurate distribution of attention and avoids the bias caused by uninterested factors such as the individual differences of operators.

Following [17], [18], we apply a quantitative metric named the Index of Pupillary Activity (IPA) to reflect the cognitive load of operators, which measures the frequency of pupil diameter oscillation based on the wavelet decomposition. Specifically, the Discrete Wavelet Transform (DWT) is performed on the pupil diameter signal to conduct analysis at multiple resolution levels. Let n-length discrete function $x^j(t) = x^j_\phi(1), ..., x^j_\phi(n)$ represent the signal at j^{th} resolution level, then the wavelet coefficients can be given:

$$x^{j-1}_\phi(t) = \sum_i h_i x^j_\phi(2t+i), \quad x^{j-1}_\psi(t) = \sum_i g_i x^j_\phi(2t+i)$$

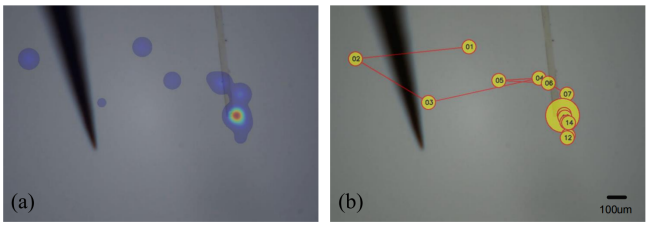

Fig. 2. A heatmap representing the concentration of the operator's gaze. A hotter color implies more concentrated attention on this point. Red is the area where the operator is most concentrated. Gaze trajectories are also shown, with numbered circles as gaze points connected in order. A circle with a larger radius implies a longer gazing time.

where $\{h_i\}$ and $\{g_i\}$ are low and high pass wavelet filters respectively. The local maxima points of coefficients $x^{j-1}_\psi(t)$ are selected and filtered by a threshold defined as $\lambda = \kappa\hat{\sigma}\sqrt{2\log n}$, where $\hat{\sigma}$ represents the standard deviation of the coefficients, and κ is chosen as 0.5. IPA is calculated as the frequency of the remaining coefficients per second, as shown in Fig. 3 (ii). We implement the DWT with PyWavelet module in Python. The wavelet function we use is symlet-8, and the decomposition level is 2.

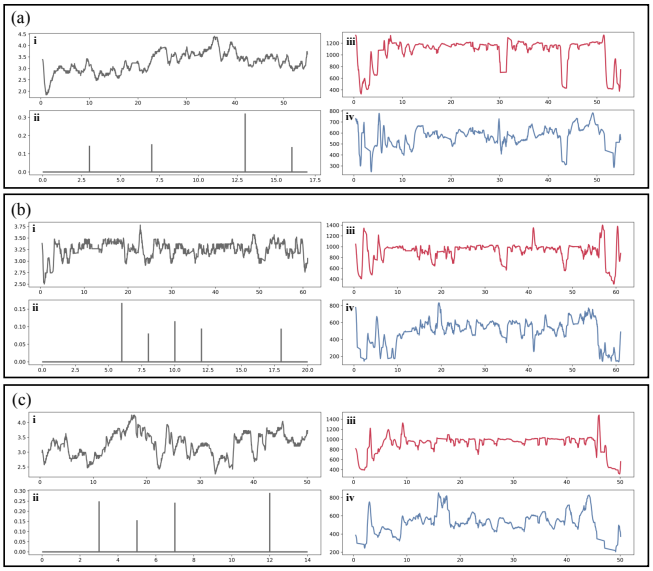

Fig. 3. Illustration of the eye movement features. (i) Pupil diameter, (ii) IPA, and (iii& iv) gaze points x & y coordinates' waveforms under motion strategies (a) A, (b) B, and (c) C of Operator 4.

B. ROI Network

The precise positions of the probe tip and the sensor's hole are required to perform the assembly task automatically. However, the target hole at the sensor tip on the image of 4096×3000 is only presented as a hole with a diameter of less than 40 pixels, which is difficult to predict accurately. The pre-mentioned gaze analysis revealed that the most concentrated areas of human attention during the assembly task are the position of the sensor tip and its vicinity. To this end, we proposed an end-to-end ROINet to mimic the human attention mechanism indicated by the gaze analysis.

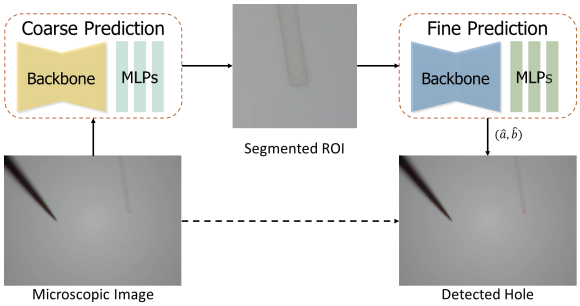

Fig. 4. Overview of the proposed ROINet. The black arrow indicates the direction of the information flow. The coarse prediction module is optimized by the gaze attention information and used to segment the ROI of the tip of the sensor. The fine prediction module is then used to predict the accurate position of the engaging hole.

The workflow of ROINet is shown in Fig. 4. The coarse prediction module first crops a small ROI (518×518) that covers the tip of the sensor, and then the ROI is sent to the fine prediction module for accurate hole position regression. We utilized a recently proposed pre-trained vision transformer Dinov2 [25] as the backbone of both modules.

The network is trained in two stages. We collect and manually annotate 14k images of the engaging hole using the microscopic system to train the network. Both the training stage is optimized with the following objective function:

$$\begin{aligned}\mathcal{L}_i =& \beta_1 \parallel \hat{P}_i(\hat{a}, \hat{b}) - P_i(a, b) \parallel_2 + \\ & \beta_2 \parallel (\hat{a} - \hat{b}) - (a - b) \parallel_2, \quad i = 1, 2\end{aligned}$$

where the $\hat{P}_i(\hat{a}, \hat{b})$ and $P_i(a, b)$ denote the prediction and ground truth hole position in i^{th} stage, respectively. The second term is a regularization loss to prevent the local minimum that the predicted hole only slides on the diagonal of the input image. The targeted point P_1 denotes the center of the selected ROI for the first stage, while the P_2 denotes the accurate position of the engaging hole in the ROI coordinate. We implemented the network with Pytorch and used Adam optimizer with an initial learning rate of 0.001. All the networks are trained on a desktop workstation with an NVIDIA RTX A6000 GPU.

IV. Experiments and Results

In this study, ten operators were recruited to perform the same micro-manipulation task with three motion strategies under teleoperated mode. Operators 1, 2 & 3 are experienced users, whereas others are novices. Task completion is evaluated based on task success, time spent, and attention distribution. A comparative analysis of the assembly process under these three motion strategies is conducted to screen and obtain a strategy with the best comprehensive results. In addition, different kinds of motion strategies are applied to the automatic mode guided by visual servoing to further demonstrate the improvement of the task after the optimized motion strategy.

A. Results of Eye Movement Features

The waveform of the original data of gaze and the IPA are shown in Fig. 3, in which we illustrate the experiment results

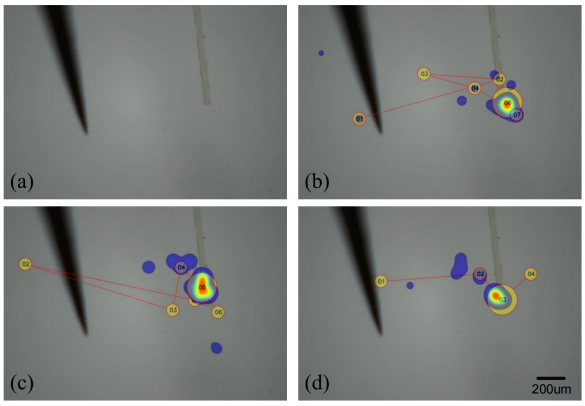

Fig. 5. Results of the gaze heatmaps and trajectories. (a) An original screenshot and heatmaps with gaze trajectories of strategies (b) A, (c) B, and (d) C, respectively.

of Operator 4 under three strategies as an example. The personalities and emotions of operators vary greatly, and such individual subjective factors interfere with the evaluation. Therefore, as a physiological indicator of mental stress, IPA is used to conduct horizontal comparisons between different strategies of the same operator. Comparative analysis of IPA between other operators was not performed. As shown in Table II, IPA does not offer any general trend, which proves that the operator's mental stress is independent of the motion strategy ($P > 0.05$). Therefore, the objectivity of our experimental results is guaranteed, and it is not affected by the individual mental bias of the operators.

As shown in Fig. 5, the results of heatmaps are clearly distinguishable. In Fig. 5(b), the gaze points are highly concentrated near the hole, which is in line with the logic of Strategy A since keeping the vertical coincidence of the probe tip and the hole at all times is the main task. The heatmap peaks in Fig. 5(c) are scattered since the operator has to pay attention to the relative position of the objectives instead of only one point. The relative relationship between the probe and the sensor is not clear. The operator is hesitant and subconsciously checks back. In strategy C, after the "palpation" confirms the successful contact of the targets, the inspection-free draws the operator's attention to a distinct single concentrated heat peak. Heatmap results imply the operator needs to pay attention to less information in Strategies A & C than in B, which is conducive to ROI segmentation and effective information screening.

B. Results of ROINet

To show the effectiveness of our proposed ROINet, we compare it with two baseline methods (Baseline1 and Baseline2) with two metrics, i.e., mean hole position error (MPE) and accuracy (Acc). Both the baselines have the same structure and training strategy as the fine prediction module, except the input images are directly resized to 518×518 instead of cropped for the Baseline1 while the input is cropped from a static region of the microscopic image for the Baseline2. The experimental results are presented in Table I, ROINet significantly outperforms all the baseline methods in all comparing metrics. Since the engaging hole

TABLE I

COMPARISON BETWEEN ROINET AND DIRECT REGRESSION METHOD

Method	Baseline1	Baseline2	ROINet
Mean Hole Error	$22.2\mu m$	$18.5\mu m$	$8.5\mu m$
Accuracy	25.7%	93.7%	95.3%

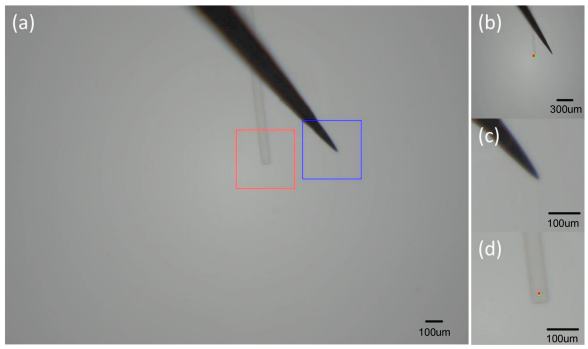

Fig. 6. The qualitative comparison between Baseline1 (b), Baseline2 (c), and ROINet (d). The red and blue rectangle denotes the segmented ROI predicted by ROINet and the static ROI for Baseline2 while the red points and green points are predicted engaging holes and the groundtruth, respectively. We adjust the brightness and contrast of the images with PowerPoint for better illustration.

has a radius of less than $12.5\mu m$, we regard a prediction of MPE $< 12.5\mu m$ as an accurate prediction. Our proposed ROINet achieves an accuracy of 95.3% with an MPE of $8.5\mu m$, superior to the Baseline2 which reaches an accuracy of 93.7% with an $18.5\mu m$ MPE. The Baseline1 performs worst, mainly because the resize operation from 518×518 to the original size could amplify the prediction error.

Fig. 6 shows the qualitative comparison between our proposed ROINet with two baselines. Since the sensor is floating on the water surface, the tip of the sensor is unstable and easily produces sudden large movements. We present the moment that the sensor suddenly moves to the left due to the probe failing to pass its tip into the engaging hole and dive into the water. The Baseline1 performs barely satisfactory, better than the Baseline2 which completely lost the engaging hole, while the ROINet handles the situation easily.

C. Results of Teleoperated Mode by Human Operators

The results of teleoperated mode by human operators are shown in Table II. The evaluation indicators mainly consider time and physiological state-related indicators, namely the average and standard deviation of time spent, success and excellence rate, total reset time, and the IPA. Tasks completed within 40 seconds and 90 seconds were defined as "success" and "excellence," respectively. A "Reset" is a process of adjustment after a failure, such as a descent overshoot or slip out of the hole. Since we allow the operator to keep trying until successful in a single task, it is necessary to record and account for the number of resets.

Regarding the average time spent on tasks, there is little difference between skilled operators and novices. While Operators 4B, 6A, 7A, and 9A struggled relatively, demonstrating average operation times longer than 80 seconds, only Operator 6A had a success rate below 80%. The description of the background (familiarity with the system) did not feed back into the task completion. Therefore, we can preliminarily admit that the data of the teleoperation mode is of reference value. For strategies A & B, the individual deviations of time are apparent without any universal trend. But all strategy C tests showed a shorter time consumption than A and B. The standard deviation (STD) of time spent is a quantitative reflection of stability and repeatability. There is still no general rule between strategies A and B. The STD of strategy C is generally lower than 15s and achieves better results than A and B in most operators. Except for Operator 5&10, the STDs of these two players under strategy C are 1.4s and 4.1s higher than A, respectively. However, the STDs under all strategies of the two are below 18s, which proves that their individual characteristics lead to general stability in operations with less influence from the strategies. Therefore, from the perspective of time-related parameters, strategy C is a more time-saving and stable choice.

Excellence rates and total resets suggest similar conclusions. Nine operators can meet the requirement of "excellence" with a higher probability in strategy C, while only five and six operators in A and B achieve at least one excellent operation, and the rates do not exceed 20% and 60%, respectively. Total resets correlate strongly with other indicators. The average time and STD are greater than 49s and 10s for tests where reset exists. Intuitively, the operator developed anxious emotions during the task or made wrong judgments about the positional relationships of experimental components, leading to resets. The total amount of resets in Strategy C is far lower than those of Strategy A & B.

Fig. 7 shows the probe tip trajectories of representative tasks of Strategy A, B, and C. One reset exists in each presented result. As shown in Fig. 7(a), the time-consuming part is fine-tuning in Strategy A, shown as many trajectory points with a broad period near the hole. Resetting in Strategy A is a simple probe raising without any detouring, which reduces both time and distance costs. The distribution density of coordinate points in Fig. 7(b) is relatively small because the operator no longer needs frequent compensations for position calibration. But the reset comes at a severe cost, requiring a detour and redoing all the maneuvers except the approaching. Strategy C's trajectory map is smoother and gentler as Fig. 7(c). The densely populated regions result from the reciprocating motion of "palpation". The reset for this test occurred during the first "palpation" when the sensor slipped due to weak contact, which was quickly corrected upon realization with little wasted time. The second "palpation" returned positive results, so here's no hesitation.

Almost all indicators and trajectory maps imply the same conclusion that strategy C is the optimal locomotion strategy for this task. In the remote operation mode, the assembly task can be completed stably in a smoother and more relaxed way within a shorter time.

D. Results of Automatic Mode with Skill Learning

To verify whether strategy C can also be applied to the automatic mode, we programmed the "palpation"-based strategy into the neural network-based visual servoing system.

TABLE II

TASK PERFORMANCE OF TEN OPERATORS USING DIFFERENT MOTION STRATEGIES DURING THE TELEOPERATED MICRO-MANIPULATION TASK

	Time spent			Success rate			Excellence rate			Total resets			Average IPA		
Name \ Strategy	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Operator 1	63.2±33.4s	55.2±28.6s	35.4±8.6s	80%	80%	80%	20%	60%	60%	5	4	1	0.125	0.123	0.122
Operator 2	52.4±10.5s	58.0±16.3s	39.4±3.5s	100%	100%	100%	20%	0%	60%	2	5	0	0.120	0.115	0.148
Operator 3	48.4±7.1s	49.6±11.8s	36.0±4.6s	100%	100%	100%	20%	20%	80%	1	2	0	0.097	0.147	0.148
Operator 4	66.2±9.2s	82.8±40.0s	47.0±7.9s	100%	80%	100%	0%	0%	20%	0	1	0	0.093	0.093	0.128
Operator 5	57.2±4.5s	72.4±14.8s	53.6±5.9s	100%	80%	100%	0%	0%	0%	0	1	0	0.131	0.083	0.121
Operator 6	142.0±52.7s	61.8±16.6s	53.6±8.2s	20%	100%	100%	0%	20%	20%	7	0	0	0.084	0.094	0.110
Operator 7	80.0±58.4s	57.2±19.5s	38.6±10.2s	80%	80%	100%	20%	40%	40%	4	2	0	0.139	0.167	0.126
Operator 8	53.2±12.7s	57.8±15.4s	36.6±4.3s	100%	100%	100%	20%	0%	80%	1	5	0	0.111	0.121	0.125
Operator 9	83.2±31.0s	68.6±35.4s	50.6±14.8s	80%	80%	100%	0%	20%	40%	2	5	1	0.096	0.098	0.172
Operator 10	49.8±4.7s	72.4±17.9s	48.8±8.8s	100%	80%	100%	0%	0%	20%	0	3	1	0.129	0.107	0.128
Auto	-	-	34.8±5.9s	-	-	90%	-	-	80%	-	-	0	-	-	-

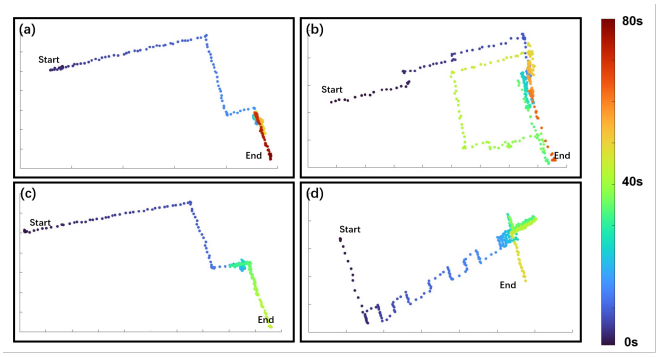

Fig. 7. Probe tip trajectories for Strategies (a) A, (b) B, (c) C, and (d) auto. Hotter colors imply tasks that have longer time costs.

Under the guidance of feature recognition, the approaching part can be completed automatically. The subsequent descent, compensation, and passing require indirect contact judgment based on "palpation". We will not give the program unlimited trial time because problems such as target loss are complex for the system to solve independently, and it makes no sense to indulge the program in trying. Therefore, irreversible errors are recorded as failures.

Out of the ten tests in automatic mode, only one failed. The results of the nine successes are shown in Table. II. To cooperate with the visual servo program and provide more obvious features, the reciprocating motion range of "palpation" in automatic mode is designed to be larger. A single "palpation" cycle takes six seconds, which is a drag on time spent. Even so, the average time and excellence rate in automatic mode are still not inferior to any operator. This proves that our attempt to migrate the motion strategy C to automatic mode is successful.

The trajectory map of automatic mode in Fig. 7 (d) reveals the difference between automatic and remote operation modes. The trajectory points are not scattered during the mission, so there are little overshoots. When approaching, the square oscillation is caused by the fluctuation of the detection results of the hole target point and has nothing to do with the strategy. Therefore, although some subtasks are intentionally prolonged in the automatic mode, hesitation and overshoot, which are familiar to human operators, are still successfully avoided. After being applied to robotic systems, strategies learned from human operating patterns are optimized again.

V. CONCLUSIONS AND DISCUSSION

The main contribution of this paper is to provide information filtering capabilities and motion strategies learned from human operators for micro-manipulation vision systems with missing information and small feature sizes. The neural network used to segment the ROI area is trained by the gaze point distribution of human operators, which simplifies information processing and reduces interference during the feature extraction. The characteristics of three feasible motion strategies were verified through human operators' testing. Among them, the strategy based on "palpation" has achieved advantages in many aspects, namely time consumption, reset time, success rate, excellence rate, and attention distribution. After applying the "palpation" strategy to the automatic mode, the depth information was successfully obtained indirectly through the relative motion relationship, and an efficient result with an average task time of only 34.8s was achieved.

It is worth noting that the proposed ROINet and "palpation" strategies are generalizable and have general values for similar micro-operation tasks. When the task changes, the idea of using an eye tracker to collect data and train with the same network structure still applies. Similarly, the idea of "palpation" can also be a reference for other micro-manipulation tasks that allow contact with a limited viewing angle.

In addition, "palpation" is a strategy that can be realized and summarized by the operator and observer, particularly for delicate micro-manipulation tasks. We believe some potential subconscious operations and features are still not perceived yet. However, it is unknown whether they can be embodied as a quantifiable strategy for the robot vision system. Some features, such as the depression of the liquid surface shown as a faint halo, can be recognized and distinguished by humans, but its detection is challenging to be satisfactory. The follow-up of this work is mainly to find more concrete human operating characteristics.

REFERENCES

[1] Y. Ma, K. Du, D. Zhou, J. Zhang, X. Liu, and D. Xu, "Automatic precision robot assembly system with microscopic vision and force sensor," *International Journal of Advanced Robotic Systems*, vol. 16, no. 3, p. 1729881419851619, 2019.

[2] Y. Wei and Q. Xu, "Design and testing of a new cell microinjector with embedded soft force sensor," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1913–1918.

[3] G.-Z. Yang, J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield *et al.*, "The grand challenges of science robotics," *Science robotics*, vol. 3, no. 14, p. eaar7650, 2018.

[4] Y. Guo, W. Chen, J. Zhao, and G.-Z. Yang, "Medical robotics: opportunities in china," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 361–383, 2022.

[5] Y. Guo, Y. Li, and Z. Shao, "Rrv: A spatiotemporal descriptor for rigid body motion recognition," *IEEE transactions on cybernetics*, vol. 48, no. 5, pp. 1513–1525, 2017.

[6] S. Liu, D. Xu, D. Zhang, and Z. Zhang, "High precision automatic assembly based on microscopic vision and force information," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 382–393, 2014.

[7] D. Zhang, Y. Ren, A. Barbot, F. Seichepine, B. Lo, Z.-C. Ma, and G.-Z. Yang, "Fabrication and optical manipulation of micro-robots for biomedical applications," *Matter*, vol. 5, no. 10, pp. 3135–3160, 2022.

[8] D. A. Duque, F. A. Prieto, and J. G. Hoyos, "Trajectory generation for robotic assembly operations using learning by demonstration," *Robotics and Computer-Integrated Manufacturing*, vol. 57, pp. 292–302, 2019.

[9] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.

[10] W. Wang, Y. Chen, R. Li, and Y. Jia, "Learning and comfort in human–robot interaction: A review," *Applied Sciences*, vol. 9, no. 23, p. 5152, 2019.

[11] E. Musk *et al.*, "An integrated brain-machine interface platform with thousands of channels," *Journal of medical Internet research*, vol. 21, no. 10, p. e16194, 2019.

[12] F. Qin, D. Xu, D. Zhang, W. Pei, X. Han, and S. Yu, "Automated hooking of biomedical microelectrode guided by intelligent microscopic vision," *IEEE/ASME Transactions on Mechatronics*, 2023.

[13] T. Tien, P. H. Pucher, M. H. Sodergren, K. Sriskandarajah, G.-Z. Yang, and A. Darzi, "Eye tracking for skills assessment and training: a systematic review," *journal of surgical research*, vol. 191, no. 1, pp. 169–178, 2014.

[14] T. Tien, P. H. Pucher, M. H. Sodergren, *et al.*, "Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair," *Surgical endoscopy*, vol. 29, pp. 405–413, 2015.

[15] G. Gras and G.-Z. Yang, "Intention recognition for gaze controlled robotic minimally invasive laser ablation," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2431–2437.

[16] K. Fujii, G. Gras, A. Salerno, and G.-Z. Yang, "Gaze gesture based human robot interaction for laparoscopic surgery," *Medical image analysis*, vol. 44, pp. 196–214, 2018.

[17] A. T. Duchowski, K. Krejtz, I. Krejtz, C. Biele, A. Niedzielska, P. Kiefer, M. Raubal, and I. Giannopoulos, "The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1–13. [Online]. Available: https://doi.org/10.1145/3173574.3173856

[18] Y. Guo, D. Freer, F. Deligianni, and G.-Z. Yang, "Eye-tracking for performance evaluation and workload estimation in space telerobotic training," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 1, pp. 1–11, 2021.

[19] C. M. Yeum, J. Choi, and S. J. Dyke, "Automated region-of-interest localization and classification for vision-based visual assessment of civil infrastructure," *Structural Health Monitoring*, vol. 18, no. 3, pp. 675–689, 2019.

[20] H. Wang, X. Lou, Y. Cai, Y. Li, L. Chen *et al.*, "Real-time vehicle detection algorithm based on vision and lidar point cloud fusion," *Journal of Sensors*, vol. 2019, 2019.

[21] S. Gündoğdu, Ö. H. Çolak, E. A. Doğan, E. Gülbetekin, and Ö. Polat, "Assessment of mental fatigue and stress on electronic sport players with data fusion," *Medical & Biological Engineering & Computing*, vol. 59, no. 9, pp. 1691–1707, 2021.

[22] C. Hirt, M. Eckard, and A. Kunz, "Stress generation and non-intrusive measurement in virtual environments using eye tracking," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 5977–5989, 2020.

[23] S. Agrawal and S. Peeta, "Evaluating the impacts of situational awareness and mental stress on takeover performance under conditional automation," *Transportation research part F: traffic psychology and behaviour*, vol. 83, pp. 210–225, 2021.

[24] Y. An, J. Yang, B. He, Y. Liu, Y. Guo, and G.-Z. Yang, "A microscopic vision-based robotic system for floating electrode assembly," *IEEE/ASME Transactions on Mechatronics*, pp. 1–11, 2024.

[25] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.