Eye-tracking for Performance Evaluation and Workload Estimation in Space Telerobotic Training

Yao Guo, *Member, IEEE*, Daniel Freer, Fani Deligianni, Guang-Zhong Yang, *Fellow, IEEE*

***Abstract*—Monitoring the mental workload of operators is of paramount importance in space telerobotic training and other teleoperation tasks. Instead of the estimation of task-specific workload, this paper aimed at investigating the impact of two significant confounding factors (time-pressure and latency) on space teleoperation and explored the use of eye-tracking technology for factor-induced mental workload estimation and performance evaluation. Ten subjects teleoperated a Canadarm2 robot to complete a complex on-orbit assembly task in our photo-realistic training simulator while wearing a head-mounted eye-tracker. To understand how time-pressure and latency influence eye-tracking features, we first performed the statistical analysis on various features with respect to a single factor and across multiple groups. Next, eye-tracking features extracted from segment data and trial data were used to identify the mental workload induced by confounding factors, which can be used for developing personalized training programs and guaranteeing safe teleoperation. Furthermore, to improve the recognition performance using segment data, we proposed the activity ratio and time ratio to characterize the informative segments. Finally, the relationship between simulator-defined performance measures and eye-tracking features was examined. Results showed that fixation duration, saccade frequency and duration, pupil diameter, and index of pupillary activity are significant features that can be used in both factor-induced mental workload estimation and task performance evaluation.**

***Index Terms*—Mental workload; space telerobotic training; eye-tracking; confounding factors; teleoperation performance**

I. Introduction

In space teleoperation, human operators need to perform complex On-Orbit Operations (O^3) (e.g., on-orbit assembly and active debris removal) from the International Space Station (ISS) or from the Earth control station [1]. Therefore, it is a challenging task to mentally determine the 3D position of the Canadarm2 robot with relation to the ISS from limited 2D visual feedback (three on-board cameras as shown in Fig. 1) while avoiding obstacles such as debris or parts of ISS [2]. Moreover, space teleoperation is particularly challenging as astronauts would suffer various extreme environmental/confounding factors.

Although many studies focused on the characterization of the workload raised by tasks with different difficulty levels [3], [4], there have been increasing studies on the mental workload induced by the presence of external confounding factors [5], [6]. Latency is one of the most significant factors that influences the operators' mental workload and teleoperation performance. In space teleoperation, latency is caused by long-range and low bandwidth communication [7], resulting in a time delay between the input command from the operator and the corresponding movement of the robot. Additionally, operators' mental workload can be influenced by varying temporal demands, i.e., the task with time-pressure [6]. In reality, the failure of astronauts to handle these situations and be resilient to stress can result in accidents. Hence, realistic simulators are human-in-the-loop systems that could provide insight into the enhancement of training protocols and operator performance [8]. Specifically, monitoring operators' mental workload induced by the external confounding factors and evaluating their training performance can be used to improve the safety in space teleoperation and develop personalized training programs [9].

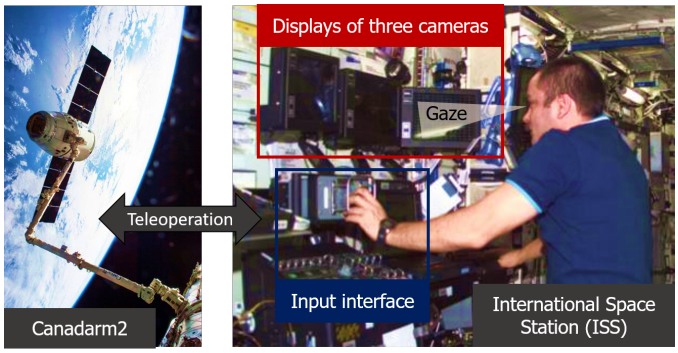


Fig. 1. An example of the real space teleoperation in the ISS. Astronauts need to control Canadarm2 robot through the input interface by observing limited visual feedback. This operation can also be performed from Earth control.

Mental workload estimation from eye-tracking data has gained increasing popularity for those needing to perform complex tasks, such as pilots [10], drivers [11], and surgeons [12]. Eye-tracking data can provide information on where and what the operator looks at, how long the operator looks at it, and which eye movements occur. Such characterization of visual interaction with complex user interfaces could help to infer the shifts of attention, fatigue, and effort of the operator during teleoperation. Moreover, unlike EEG and other sensing modalities [13], [14], eye-tracker is less interfered by cosmic radiation in space. However, there is little effort for eye-tracking based factor-induced workload estimation.

The objective of this paper is to investigate eye-tracking for factor-induced mental workload estimation and performance evaluation during space telerobotic training. We developed

This work was supported by Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/R026092/1.

Y. Guo and G.-Z. Yang are with Institute of Medical Robotics, School of Biomeidical Engineering, Shanghai Jiaotong University. D. Freer is with The Hamlyn Centre for Robotic Surgery, Imperial College London. F. Deligianni is with School of Computing Science, University of Glasgow.

a photo-realistic O^3 teleoperation simulator [8], which is able to modulate two common yet significant confounding factors, i.e., **latency** and **time-pressure**. Within-group experiments were carried out by ten subjects using the simulator under multiple conditions while their eye movements were recorded using a wearable eye-tracker. Various eye-tracking features were analyzed, including eye movement features, eye blink, pupillary response, and Point of Gaze (PoG) on different cameras. We first performed the Analysis of Variance (ANOVA) to investigate how different features will change with respect to the induction of time-pressure and latency. Different from the two-class classification on low or high task-specific workload as in [15], [16], this paper focused on factor-induced workload identification from eye-tracking data. Both two-class recognition on the presence of a single factor and multi-class factor-induced workload classification were conducted by using various eye-tracking features extracted from the segment and trial data, respectively. As workload estimation from segment data can be used to provide real-time feedback, we further proposed the *activity ratio* and *time ratio* strategies to distill the informative segments for training and testing, thus improving the recognition performance. The activity ratio represents the level of interaction with the system and the time ratio characterizes how the workload changed over time. Furthermore, to explore eye-tracking for performance evaluation, we analyzed the correlation between eye-tracking features and simulator-defined performance measures.

The main contribution of this paper are three-fold:

- The impact of two key confounding factors (time-pressure and latency) on operators' workload in space telerobotic training is investigated through eye-tracking data.
- To the best of our knowledge, this is the first work that investigates the effect of latency on eye movements and explores factor-induced workload identification via eye-tracking data.
- The informative segments via the activity ratio and time ratio strategies are extracted to improve the recognition performance.

This paper is organized as follows. Related literature is discussed in Section II. Section III describes the O^3 training simulator and the wearable eye-tracker. In Section IV, we introduce the training dataset, data normalization methods and eye-tracking features. Experimental results on statistical analysis, factor-induced workload identification, and performance evaluation using eye-tracking data are shown in Section V. Section VI concludes the results and discusses possibilities offered by the developed techniques.

II. Related Works

A. Latency and Time-pressure in Teleoperation

Previous studies have reported that task performance was degraded and NASA-TLX scores were increased in various teleoperation scenarios with latency [5], [17]–[19]. In [17], four control latencies ranging from $0s$ to $4s$ were added to the rotational and translational motion control. When the latency is above $1s$, people will change their control strategies from continuous control to "move and wait". Khasawneh *et al.* [18] investigated the relationship between $500ms$ latency and operators' performance in teleoperating a rescue robot with different levels of complexity. Lu *et al.* [19] evaluated the impact of $800ms$ time delay in dual-task teleoperation of unmanned ground vehicles. In addition, Yang *et al.* [5] showed that operators' frustration, anger, and workload were increased by adding time delay to a robotic navigation task.

Recent studies have shown that adding time pressure has a significant effect on surgeons' mental states in teleoperated robotic surgery [6], [20]. Singh *et al.* [6] reported that surgeons had an increase in performance with moderate temporal stress, followed by degradation with a further increase of time-pressure. Results in [20] showed that compared to novice groups, senior surgeons coped better with time pressure and exhibited greater performance on stability during a laparoscopic suturing task. In a recent study on a driving simulator, Rendon *et al.* [21] investigated the effects of time pressure on measures of eye-tracking data, task performance, and other physiological signals. Results showed that under time pressure, operators finished the task more efficiently and exhibited a significant increase in respiration rate, heart rate, and pupil diameter, and reduced blink frequency. However, there is a lack of studies on the effect of multiple confounding factors.

B. Workload Estimation from Eye Tracking Data

In terms of eye-tracking data, the most widely used features indicating workload metrics include the pupillary response, eye fixations, eye saccades, and eye blinks [22], [23]. In the 1960s, early researches have revealed that the pupil diameter increased with the task difficulty [24], which is highly related to the mental workload. However, the pupillary response is sensitive to the changes of luminance or emotional arousal [25]. To reduce the influence of luminance and emotional arousal from the pupillary response, Marshall *et al.* [26] found that the cognitive workload can be estimated more accurately from small rapid pupil dilations and proposed an Index of Cognitive Activity (IOCA) metric. Furthermore, an improved Index of Pupillary Activity (IPA) [27] was proposed by performing wavelet analysis to identify the abrupt pupil dilations from the diameter sequences. Higher IOCA/IPA values indicate a higher degree of cognitive workload during the task [26], [27]. In terms of eye movements, eye fixations and saccades were extensively used for mental workload estimation [28]. Eye fixations indicate that gaze point is maintained on a small region for a considerable time interval, whereas eye saccades represent rapid gaze movements from one region to another. Prior studies observed that the fixation duration has a negative relationship with mental workload [10], [29]. No significant relationship between the speed/amount of saccades and mental workload has been observed in [10], [29], [30]. In [22], [30], researchers found that blink frequency and blink duration exhibited a positive relationship with the mental workload. Benedetto *et al.* [11] showed that blink duration, compared to blink frequency, is more sensitive and reliable for workload detection.

Workload identification/recognition is another important and yet not well-studied topic. Most existing works mainly explored the capability of workload recognition from EEG data

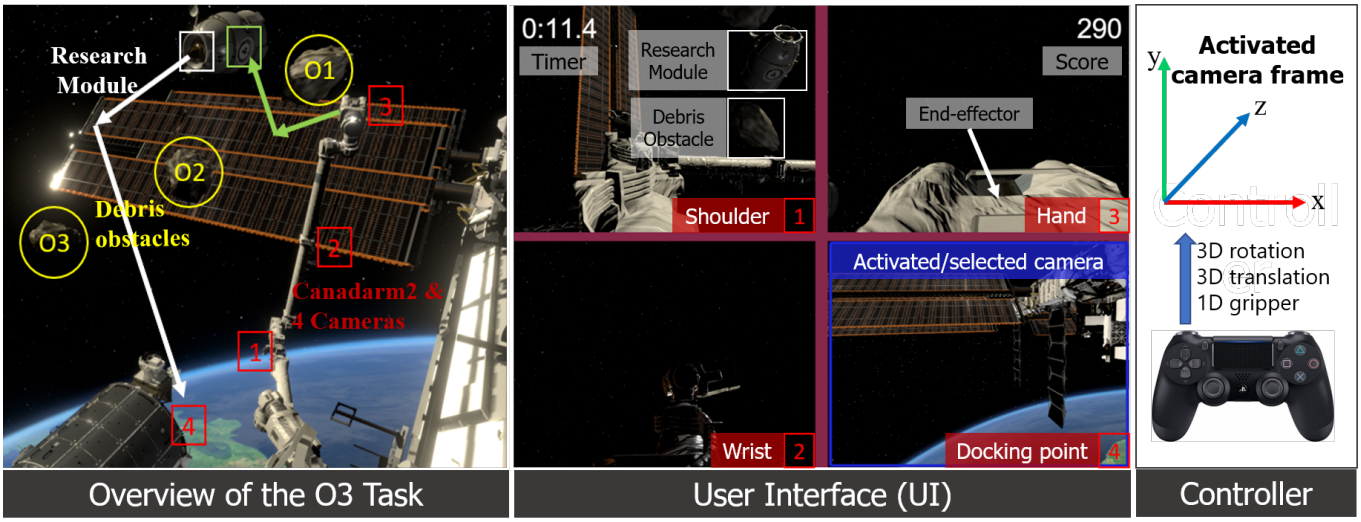


Fig. 2. The left image demonstrates an overview of our O^3 simulator from a global view. The middle part shows the user interface observed by operators. Cameras #1-#3 are located at the shoulder, wrist, and hand of Canadarm2, and camera #4 is near the docking point. The timer is shown in the top left and the performance score is displayed in the top right. The research module and the debris can be found in camera #1. Besides, the activated camera is highlighted with a blue box. The right part shows the controller and the control strategy.

[13], [31] or the combination of EEG with other physiological signals [14], [32]. Most recently, workload recognition from eye-tracking data has attracted increasing attention. Wu *et al.* [15] utilized eye-tracking features to predict two workload levels (low & high) characterized by NASA-TLX scores. In [16], Hecht *et al.* modeled the eye gaze behavior as the tradeoff between a non-goal oriented saliency distribution and a reward task-related distribution, and performed the classification of high or low workload by using these two gaze patterns. Nevertheless, there is few research focusing on the identification of the workload raised by external confounding factors.

C. Eye-tacking for Performance Evaluation

Eye-tracking data can serve as an objective tool for performance evaluation and skill assessment [23]. Recent works have reported that eye movements were significantly different between novice and expert groups of pilots [10], nurses [33], and surgeons [34], [35]. This could provide evidence of using eye-tracking in evaluating training performance. Ziv *et al.* [10] found that expert pilots had more fixations on different instruments and shorter dwell time on each instrument compared to novices. In [33], eye fixations and PoG on the area of interests were used to predict the performance score of both qualified and training nurse groups. For surgeons, differences in eye metrics reflecting focused attention can also be found between junior and senior surgeons [34], [35]. As senior surgeons did not experience the same degree of cognitive workload, they showed higher fixation frequency, and simultaneously, lower IOCA values over junior surgeons. Besides, Hamada *et al.* investigated the estimation of drivers' distraction from eye saccade movement [36]. They reported that in terms of distracted driving conditions, older groups typically exerted increased mental workload.

III. System Description

A. O^3 Training Simulator

The current version of our photo-realistic O^3 training simulator simulated an assembly task by teleoperating Canadarm2 with a Playstation controller [8], where Canadarm2 is a seven Degree of freedom (DOF) robotic arm and has been widely used for completing O^3 tasks. The task consists of three key steps: 1) move the end-effector of Canadarm2 to grasp the research module; 2) move the module toward its docking point on the ISS while avoiding the collisions; 3) dock the module onto the docking point of the ISS.

As demonstrated in Fig. 2, similar to the real scenario, three cameras were deployed on the shoulder, wrist, and hand of Canadarm2 to provide 2D visual feedback to the operators. In addition to these three cameras, we added another camera near the docking point to help the operator to complete the docking accurately. During the task, operators used a controller to realize the 6D movement of the end-effector in the activated/selected camera frame, select and rotate the camera, and operate the gripper to grasp and release the module.

The presence of debris in space poses a risk to space missions, which can damage the spacecraft and the infrastructures on the ISS [2]. Hence, adding debris in the simulator can provide valuable exercise for teleoperating the robot. Before recording data, operators were first able to familiarize themselves with the simulated O^3 task without the presence of debris. During the formal training, three pieces of debris with static initial positions were added, which would only move after a collision. To quantitatively evaluate operators' teleoperation performance and provide real-time feedback, we developed a reasonable scoring protocol as introduced in [8]. During the task, the score and timer were displayed on the top left and top right as shown in the middle part of Fig. 2. For two subtasks - grasping and docking, we calculated *Grasp score* and *Dock score* by considering the connection

quality. The quality was calculated using $Q = 100/(dist + 0.1) + 5000/(\theta_{dist} + 5)$, where $dist$ is the Euclidean distance between the executed and ideal position for each subtask, and θ_{dist} is the angular distance by comparing the quaternions of the ideal and actual poses. Moreover, the *Final score* of the task was calculated by considering the elapsed time for each trial, collisions, *Grasp score* and *Dock score*. We refer the readers to [8] for more details on our O^3 training simulator and performance score calculating.

B. Workload Modulation via Confounding Factors

Most prior works analyzing workload during teleoperation focused on the estimation of the inherent workload induced by the tasks themselves [3], [4], [9]. In contrast, this paper aims to investigate the impact of external confounding factors on operators' mental workload. To achieve this, we modulated (with or without) two confounding factors in the simulator, i.e., **latency** and **time-pressure**, which enables us to examine the associated changes in operators' eye-tracking data and thus evaluating operators' mental workload and performance.

1) Latency: For O^3 tasks, astronauts need to cope with the latency originating from the long-range and low bandwidth communication link [7]. Except for teleoperating Canadarm2 from ISS, such a task sometimes is assisted by operators from the Earth control center with hundreds of milliseconds of latency [7]. In the near future, astronauts may also need to teleoperate the robot on Mars/Moon from its orbit. Hence, studying the effect of latency plays a critical role in space teleoperation. In the experiments, we set it as a ($0s$, $0.5s$ & $1.0s$) during each trial and randomized it across trials.

2) Time pressure: Most space teleoperation tasks will encounter time limitations due to power or fuel requirements. Previous studies have proven that time pressure significantly affected the mental workload of surgeons [6], especially for tasks requiring high precision. To this end, we set an option that forces the task to be completed within $4mins$, which aims at investigating the changes of operators' mental effort under temporal stress. With added time pressure, the simulator will automatically stop when it meets the time limit.

C. Wearable Eye-tracking Device

To track the operators' eye movements and pupillary responses, each subject was asked to wear a head-mounted PupilLabs Core eye-tracker during the training, which can capture the eye-tracking data more precisely than table-mounted devices [37]. As shown in Fig. 3, the eye-tracker consists of a scene camera and two eye cameras. Given the images captured from the two eye cameras, the 2D coordinates (x_p, y_p) of the pupil center and the diameter of the two pupils d_p can be estimated as image pixels. By integrating the information estimated from two eyes, the 2D Point of Gaze (PoG) can be further determined and projected onto the scene image. Furthermore, eye movement behaviors can be detected from eye images. The scene camera captures the main screen that displays the O^3 experiment. To acquire reliable 2D gaze position estimation, eye tracker calibration was performed after two trials. The frequencies of the eye cameras and the world camera are 120Hz and 30Hz@1080p, respectively.

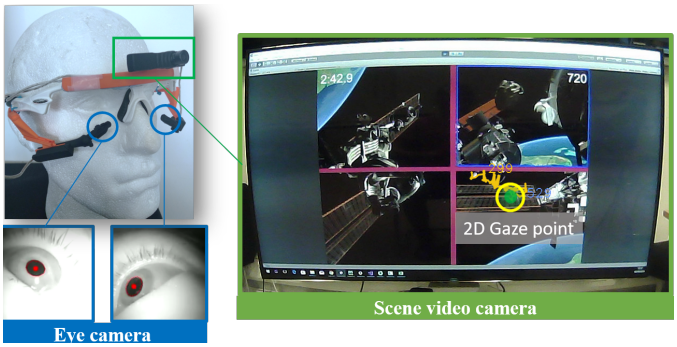


Fig. 3. The head-mounted PupilLabs eye-tracker consists of one scene video camera and two eye cameras. By fusing the information from two eyes, the point of gaze can be predicted onto the 2D scene image plane.

TABLE I
FOUR LEVELS OF FACTOR-INDUCED WORKLOAD

Category	Notation	Latency	Time pressure
Low workload	low	N	N
Latency	lat	Y	N
Time pressure	tp	N	Y
Time pressure and Latency	tp+lat	Y	Y

IV. Data Collection and Eye-tracking Features

A. Study Protocol

In the study, ten healthy subjects (two females and eight males) with normal/corrected vision and without known physical/mental problems participated in the simulated O^3 teleoperation experiments. Ethical approval for this experiment was received under No. ICREC-18IC4816, and all subjects gave informed consent. In the experiments, each subject was asked to sit in front of the visual display while wearing a PupilLabs eye-tracker. We designed the experiment with the within-group (repeated measures) protocol, which means the same participants took part in the experiments in terms of four conditions as listed in Table I, which allows us to investigate workload modulation through the combination of latency and time-pressure. Before conducting the experiment, operators were able to complete several "familiarization" trials. For the experiments without these two factors, we denoted it as 'low workload', which means that the workload was only raised by the O^3 task. The streamed data of the controller input, simulator-defined measures, and eye-tracking data were synchronized by LabStreamingLayer[1].

In this paper, we performed factor-induced workload estimation using the eye-tracking features extracted from both **trial data** and **segment data** was conducted. The segment data indicates the data within a time window of length t_w, which could be used to analyze the workload in a short time interval, which could provide real-time feedback. In order to compare the performance using segments of different lengths, we evaluated segment data with t_w= $\{2s, 5s, 10s, 20s\}$ non-overlapping windows.

[1]https://github.com/sccn/labstreaminglayer.

B. Data Normalization

Due to subject-specific differences, there exist significant variations across subjects in terms of the teleoperation and the corresponding eye-tracking data. These significant variances mainly originate from the following sources, including different pupil sizes or eye movement behaviors and the impact of other factors such as body temperature, emotion, and fatigue. Thus, we followed common normalization methods [15], [16], [31] to reduce subject differences and biases.

1) Single-trial normalization: Single-trial normalization was carried out in order to normalize the variations among trials for each subject. For instance, the pupil diameters d_p of the operator before the first trial could be smaller than that before the later trials, due to fatigue or increased mental workload resulted from previous trials. To reduce such bias, we took a segment of size t_b before the experiment to extract the baseline pupil diameters d_p^{base}, where $t_b = 10s$ for the trial data and t_b equals to t_w for the segment data. Then, the normalized pupil diameter is $\hat{d}_p = d_p - d_p^{base}$ and the pupillary response related features can be further extracted.

2) Subject-specific z-score normalization: Next, z-score normalization was adopted for calculating the normalized feature $\hat{x} = (x - \mu)/\sigma$ for each subject, where μ and σ are the mean value and standard derivation of feature x. Such normalization can compensate for individual differences and reshape the distribution of x to an approximately normal distribution, which is needed to perform the statistical analysis.

C. Eye Tracking Features

Eye-tracking data can provide informative clues to characterize visual interaction with complex user interfaces [23]. Specifically, the features indicating eye fixation, eye saccade, eye blink, gaze, and pupillary response were extracted in this paper. The detailed definitions of these features are listed in Table S-A in the supplementary material.

1) Eye movement related features: Using the detected 2D PoG, three basic eye movements can be defined: fixation, saccades, and smooth pursuits. Eye fixation is when the PoG is within a particular area or if the gaze velocity is smaller than a threshold. Eye saccade indicates the shift between fixations and can also be determined by the gaze velocity. Compared to fixation and saccade movements that have become popular for studying cognition and memory, few studies have reported the significance of smooth pursuit (eye movement that follows a moving object) in workload detection [38]. To sum up, the eye movement related features $\mathcal{F}_{eyemove} = \{\mathcal{F}_{fixation}, \mathcal{F}_{saccade}\}$ were extracted from each trial/segment, where $\mathcal{F}_{fixation}$=$\{$*Fixation frequency*, *Mean fixation duration*, *Max fixation duration*$\}$ and $\mathcal{F}_{saccade}$=$\{$*Saccade frequency*, *Mean saccade duration*, *Max saccade duration*, *Mean saccade speed*, *Max saccade speed*$\}$.

2) Blink-related features: Eye blink indicates a temporary closure of both eyes, involving movements of the upper and lower eyelids [28]. Previous studies also found that the blink frequency and blink duration have shown a correlation with respect to mental workload [11]. Inspired by this, the blink related features $\mathcal{F}_{blink}$ =$\{$*Blink frequency*, *Mean blink duration*, *Max blink duration*$\}$ were calculated.

3) Pupillary response related features: In recent decades, extensive research has revealed that pupil diameter has a positive correlation with task difficulty [30], [39]. To reduce the impact of luminance, a robust metric Index of Pupillary Activity (IPA) [27] was proposed, which can detect the rapid pupil dilations through wavelet analysis. In this paper, the pupillary response related features include $\mathcal{F}_{pupil}$ =$\{$*IPA#0*, *Mean pupil#0 diameter*, *Pupil#0 deviation*, *Max pupil#0 diameter*, *IPA#1*, *Mean pupil#1 diameter*, *Pupil#1 deviation*, *Max pupil#1 diameter*$\}$, where #0 and #1 indicate the left and right pupil, respectively.

4) PoG on different camera displays: The visual interaction with the user interface is also significant in workload estimation and performance evaluation. However, one of the inherent challenges of head-mounted devices is head movement. As 2D PoG is represented in the scene image plane, operators' head movement adds difficulty to the estimation of gaze on the visual displays. To overcome this, we detected the boundaries of the visual displays in real-time using image-processing techniques (see Fig. S-A in supplementary material). By accurately localizing the cross boundaries, the PoG on a specific camera can be determined. Accordingly, we calculated the ratio of PoG on cameras $\{$#1, #2, #3, #4$\}$ and the ratio of PoG outside the user interface within each trial as the informative features.

V. Experimental Results for Eye-tracking based Workload Estimation & Performance Evaluation

A. Statistical Analysis of Eye Tracking Features

1) Statistical analysis on a single confounding factor: To provide a dedicated view on how eye-tracking features change with a single confounding factor, we first conducted a one-way Analysis of Variance (ANOVA) on two groups (i.e., w/ and w/o a factor). The z-normalization facilitated the feature distribution across different subjects to an approximately normal distribution. Note that we mainly focused on investigating the statistical differences among different groups, and therefore, the p-values were reported. The p-value results of different features with respect to *time-pressure* and *latency* are listed in Table II. It should be pointed out that statistical analysis was only performed for trial data, as the variances of features extracted from segment data are too large. The p-value results of the simulator-based performance measures were presented in our previous study [8].

With added *time-pressure*, $\{$*Max saccade duration*, *Blink frequency*$\}$ decreased and *Max fixation duration* increased, which reveals that operators paid more attention and changed their gaze points across cameras more frequently during the task. It can also be observed that *IPA* value was significantly increased, which implies that the cognitive workload of operators under time-pressure was higher than the trials without time-pressure. In addition, we found that *Pupil deviations* decreased, which could indicate that the workload remains at a high level.

In terms of *latency*, only decrease in *Mean fixation duration* and increase in *Saccade frequency* were statistically significant, showing the difficulty in discriminating the effect of latency from eye-tracking data. According to operators' feedback, the impact of latency was obvious at the beginning of

the task. Hence, this observation could indicate that operators changed control strategies to compensate for latency.

2) Statistical analysis across four groups: Next, we conducted a one-way ANOVA of the features extracted from trial data on four workload groups as in Table I, enabling a pair-wise comparison between any two groups. For *Mean fixation duration* in Fig. 4(a), compared to *low*, we can observe a decrease in *tp* (not significant), *lat*, and *tp+lat*. In terms of IPA (combining IPAs from Eye#0 and Eye#1), there was a significant increase with *tp* compared to *low* and *lat* groups as in Fig. 4(b), which implies that added time-pressure raises the workload during teleoperation, and that added latency has less impact on pupillary response or sometimes decreases the cognitive workload features (comparison between *tp* and *tp+lat*). More important, for saccade related features (see Figs. 4(c)-(d)), a statistically significant increase can only be found in *tp+lat* compared to that in *low* and *lat* groups. Combined with the results in Figs. 4(a), it shows that operators had more fixations on different cameras and shorter dwell time on each camera under *tp+lat*.

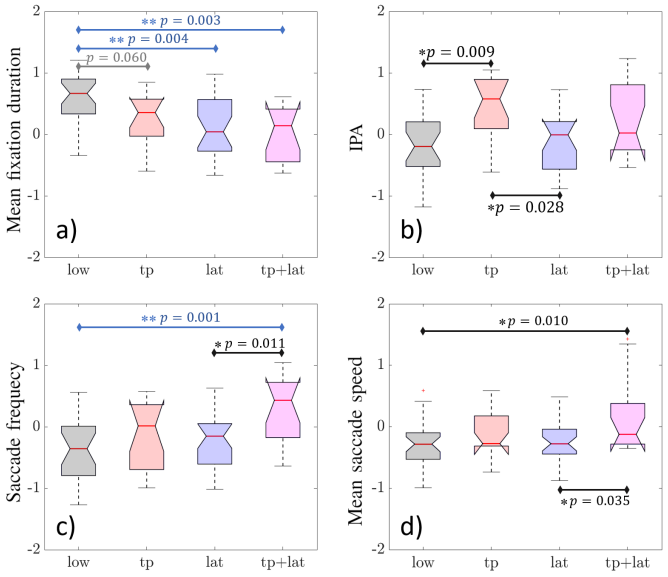


Fig. 4. ANOVA results of a pair-wise comparison among four workload levels {*low*, *tp*, *lat*, *tp+lat*}. a) Mean fixation duration; b) IPA; c) Saccade frequency; d) Mean saccade speed. $^{**}p \leq 0.005$, $^{*}0.005 < p \leq 0.05$.

B. Eye-tracking based Factor-induced Workload Identification

Unlike prior works focusing on the discrimination of low or high workload in performing the same teleoperation task [15], [16], this paper aimed to recognize the factor-induced workload from eye-tracking data, thus identifying the confounding factors experienced by operators, which could help to guarantee safe teleoperation.

1) Materials and recognition tasks: Both segment and trial data were evaluated for factor-induced workload identification in this paper. For segment data, we set the size of window $t_w = \{2s, 5s, 10s, 20s\}$ and the overlap of two consecutive segments is 0. Workload identification from segment data would be necessary to implement real-time feedback in the future. Two types of recognition tasks were performed: 1) two-class recognition by discriminating w/ or w/o a single confounding factor (*time-pressure* or *latency*); 2) multi-class classification on four levels of workload as listed in Table I.

2) Evaluation protocol and metrics: To evaluate the effectiveness of the proposed method in factor-induced workload recognition, a leave-one-subject-out (LOSO) protocol was applied to validate the generalization ability of the model to new subjects. Within the LOSO protocol, one subject was chosen for testing and the data from the remaining subjects were used for training in each run. We repeated the experiments ten times (the number of subjects) and reported the average results.

To address the imbalance problem of two-class classification, the final reported `precision` and `recall` were the average value by taking (w/ factor = positive) and (w/o factor = negative) respectively. Then we calculated the final `F1 score` to evaluate the performance. In terms of multi-class workload identification, the average 5-fold validation accuracy `(ValAcc)` during training and the test accuracy `(TestAcc)` under LOSO evaluation protocol were calculated.

3) Feature categories and classifier: In order to examine the capability of different eye-tracking features to discriminate the factor-induced workload, we examined the performance with respect to different features (i.e., *blink*, *saccade*, *fixation*, *pupil*, *eyemove*, *all*, and *ANOVA*). Among these, *blink*, *saccade*, *fixation*, and *pupil* are those listed in Table II. *eyemove* represents {*saccade*, *fixation*}, and *all* indicates the combination of {*blink*, *saccade*, *fixation*, *pupil*}. *ANOVA* represents the features with statistically significant differences with respect to *time-pressure* or *latency* as shown in Table II. Taking the feature vector extracted from each trial or a segment as input, we adopted a Support Vector Machine (SVM) with Radial Basis Function kernel as the classifier.

4) Results on two-class workload classification: Recognition results in discriminating the workload induced by adding a single factor under LOSO protocol are reported in Table

TABLE II
P-VALUES OF COMPARING FEATURES W.R.T. A SINGLE FACTOR

Category	Eye-tracking features	Time pressure	Latency
Eye movement related features	Fixation frequency	0.842	0.118
	Mean fixation duration	0.153	0.039*↓
	Max fixation duration	0.031*↑	0.164
	Saccade frequency	0.872	0.025*↑
	Mean saccade duration	0.188	0.331
	Max saccade duration	0.042*↓	0.680
	Mean saccade speed	0.463	0.335
	Max saccade speed	0.518	0.424
Blink related features	Blink frequency	0.028*↓	0.318
	Mean blink duration	0.697	0.548
	Max blink duration	0.568	0.994
Pupillary response related features (Eye#0 - Eye#1)	IPA#0	0.009**↑	0.789
	Mean pupil#0 diameter	0.255	0.548
	Pupil#0 deviation	0.047*↓	0.552
	Max pupil #0 diameter	0.268	0.917
	IPA#1	0.049*↑	0.673
	Mean pupil#1 diameter	0.474	0.909
	Pupil#1 deviation	0.039*↓	0.549
	Max pupil #1 diameter	0.556	0.854

• $^{**}p \leq .005$, $^{*}0.005 < p \leq .05$;
↑, ↓ indicate the changes of mean value $\mu_{w/}$ compared to $\mu_{w/o}$.

TABLE III
TWO-CLASS RECOGNITION RESULTS UNDER LOSO PROTOCOL

		Segment data						Trial data		
		5s			20s					
	Features	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
Time Pressure (tp)	blink	.484	.499	.492	**.525**	**.529**	**.527**	.503	.477	.488
	saccade	**.619**	.516	.563	.613	**.532**	**.570**	.515	.552	.532
	fixation	**.576**	**.548**	**.562**	.483	.501	.492	.622	.651	.636
	pupil	.526	.504	.515	**.574**	**.549**	**.561**	.604	.566	.584
	eyemove	.525	.508	.516	**.575**	**.527**	**.550**	.631	.638	.634
	all	**.547**	.498	.521	.545	**.518**	**.532**	.689	.701	.695
	ANOVA	.522	.505	.513	**.604**	**.542**	**.572**	.749	.751	.750
	Features	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
Latency (lat)	blink	.444	.466	.455	**.477**	**.483**	**.480**	.544	.560	.552
	saccade	**.584**	**.574**	**.579**	.563	.548	.555	.545	.546	.545
	fixation	.514	.515	.515	**.538**	**.538**	**.538**	.559	.548	.553
	pupil	**.560**	**.554**	**.557**	.533	.529	.531	.628	.626	.627
	eyemove	.516	.511	.513	**.545**	**.544**	**.544**	.606	.562	.583
	all	.514	.516	.515	**.539**	**.540**	**.540**	.611	.569	.589
	ANOVA	.537	.534	.535	**.557**	**.554**	**.556**	.707	.708	.707

• **Bold** indicates the highest value in each row (feature) of the segment data, and Underline is the maximum in each column. Blue background highlights the best result for recognizing each factor from the segment data and trial data.

III (results with {2s, 10s} window sizes are listed in the supplementary material). The best result on trial data was achieved by using *ANOVA* feature (F1: *tp* - 0.750, *lat* - 0.707), which are superior to those on segment data (F1: *tp* - 0.572, *lat* - 0.579). In terms of comparing segment data of different sizes, most of the best results were achieved by using features extracted from the segment of 20s. This is due to the large variations of the features extracted from small segments. It should be pointed out that for segments with smaller window sizes, *saccade* outperformed other features.

5) Results on multi-class workload classification: Recognition rates on four-level recognition task under a LOSO protocol are reported in Table IV. The best TestAcc (49.32%) was achieved by using the *ANOVA* feature extracted from each trial. As for segment data, the result using *saccade* feature from 20s segments (38.85%) outperformed others. It also can be seen that higher ValAcc was achieved with smaller segments due to the increased size of training and test sets, while better TestAcc was seen with segments of larger window size. However, for segment data, the multi-class workload classification results are not satisfactory. This is because that the main challenge in recognizing factor-induced workload from segment data is the large variance of segments originating from the intrinsic workload caused by the task itself. However, such task-specific workload is hard to be accurately modeled due to the high complexity of the O^3 task. Besides, due to the subject differences, operators may exhibit distinct eye movement behaviors and pupillary responses in terms of the same confounding factors.

C. Factor-induced Workload v.s. Task-specific Workload

Considering the challenges in recognizing factor-induced workload from segment data, we aim to offer an alternative way to improve the performance for factor-induced workload identification. Intuitively speaking, the factor-induced work-

TABLE IV
RECOGNITION ACCURACIES ON MULTI-LEVEL WORKLOAD DETECTION

Features	Segment data				Trial data	
	5s		20s			
	ValAcc	TestAcc	ValAcc	TestAcc	ValAcc	TestAcc
blink	**46.29**	34.73	42.19	35.89	32.09	33.14
saccade	**48.73**	37.47	42.59	**38.85**	35.63	34.37
fixation	**55.57**	28.52	49.62	**35.10**	37.56	37.03
pupil	**65.51**	29.36	50.59	**32.17**	32.29	36.31
eyemove	**58.85**	28.19	47.17	**37.61**	36.04	38.21
all	**65.01**	31.25	52.81	**34.53**	34.52	39.16
ANOVA	**67.20**	29.45	54.70	**33.58**	48.37	49.32

• Each element indicates the recognition accuracy (%);

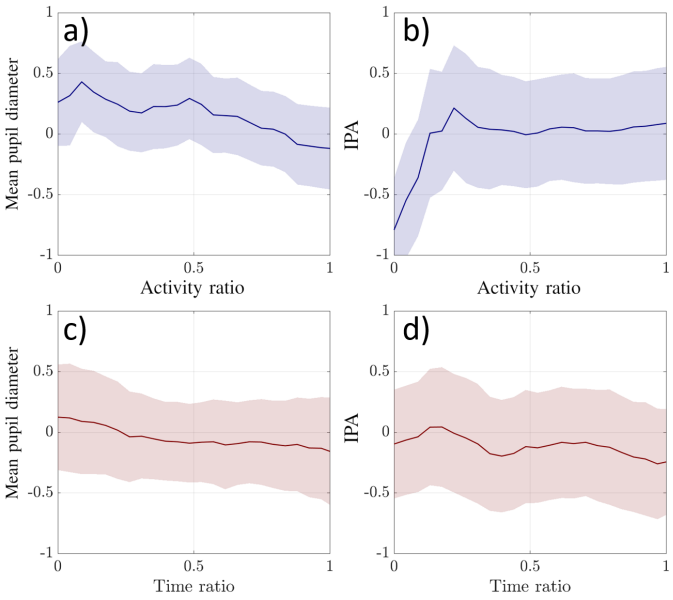
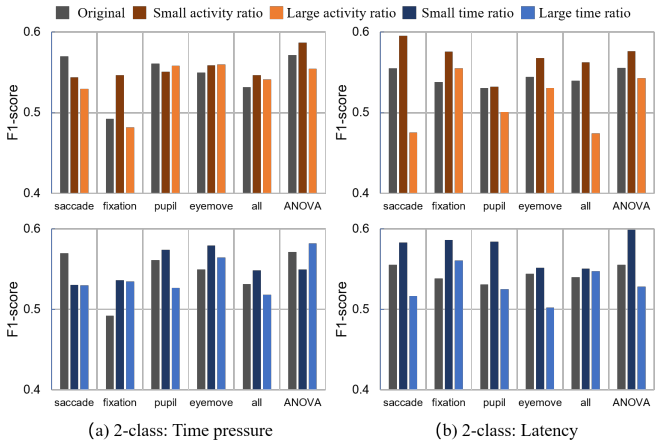
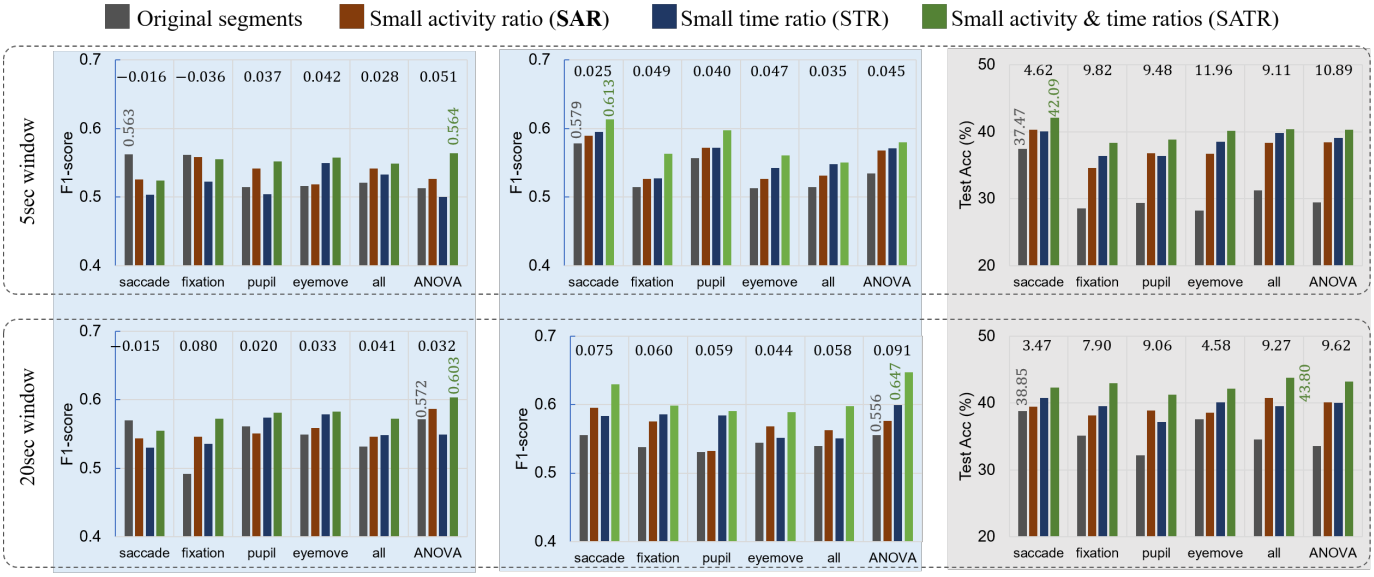


Fig. 5. Illustration of the variation of the normalized features with respect to different activity & time ratios. The solid line is the rolling average with a window of size 0.1, and the shadow area indicates the values within mean±standard derivation (std). a) Activity - mean pupil diameter; b) Activity - IPA; c) Time - mean pupil diameter; d) Time - IPA.



(a) 2-class: Time pressure (b) 2-class: Latency

Fig. 6. Comparison of results using different features extracted from 20s segment data, and the proposed configurations {SAR, LAR, STR, LTR}.



(a) 2-class: Time pressure (b) 2-class: Latency (c) Multiple levels

Fig. 7. Comparison recognition results by using different eye-tracking features extracted from segment data and the segments with the proposed configurations. Black numbers above each group of bars represent the improvement of the results by using the proposed **SATR** configuration over original segment-based results. The rotated color numbers are the best results achieved by using segment data and the proposed configurations, respectively.

load can be discriminated more easily from segments less influenced by the task-specific workload.

Without loss of generality, the task-specific workload in each segment can be assumed to be proportional to the interaction with the system. We first calculated the '**activity ratio**' $r_{act} = t_{input}/t_w \in [0, 1]$ to characterize the interaction level of a segment of size t_w, where t_{input} is the amount of time within this segment that the operator is issuing a command to the robot via the controller. In addition, considering that both factor-induced and task-specific workload will vary over time, we calculated the '**time ratio**' $r_{time} = t_{seg}/t_{trial} \in (0, 1]$ for each segment, where t_{seg} is the elapsed time from the start of the experiment to the end of the segment and t_{trial} is the total time spent for this trial. Fig. 5 plots two representative features, *Mean pupil diameter* and *IPA*, with respect to different activity and time ratios. In Figs. 5(a)(c), *Mean pupil diameters* drops with the increase of activity/time ratios. For *IPA* in Figs. 5(b)(d), it increases with small activity ratio and fluctuates with small time ratio, while remaining at a similar level with large activity/time ratios.

Inspired by these observations, we first proposed four different configurations to extract informative segments for improving factor-induced workload recognition performance, i.e., **small activity ratio (SAR)**, **large activity ratio (LAR)**, **small time ratio (STR)**, and **large time ratio (LTR)**. The threshold for determining small and large activity and time ratio was set as the mean value calculated from the corresponding training set. In this way, the informative segments can be distilled to form the new training and testing subsets through the use of 'activity ratio' and 'time ratio'. Fig. 6 shows the comparison results with the original 20s segments and the proposed four configurations. It can be seen that the **SAR** and **STR** configurations outperformed those results by **LAR**, **LTR**, and the orignial ones.

Thus, we further proposed another configuration called **small activity & time ratio (SATR)**. Fig. 7 compares the results with different features extracted from the original segment data and the segments distilled by the proposed configurations. It can be seen for *time-pressure*, **SATR** slightly outperformed original results in most cases, which could emphasize that the workload induced by time-pressure remained at a similar level with different activity and time ratios. For identifying *latency*, the proposed **SATR** configuration achieved superior results when compared to all the features over those using original segments, **SAR**, and **STR**. As shown in Fig. 7(b), `F1` of 5s window was increased from 0.576 to 0.613 (**SATR**) and `F1` of 20s achieved 0.647 (**SATR**) exceeding the second best by 0.091. This implies that the latency-induced workload was more obvious at the beginning of the task and at a low interaction level.

By comparing results on multi-class recognition as shown in Fig. 7(c), the results by **SAR** and **STR** consistently outperformed those using original segment data in most cases. It should be pointed out that the proposed **SATR** achieved the best recognition accuracies (5s, *saccade* - 42.09%; 20s, *all* - 43.80%) among five configurations, which exceeded the original results by 4.62%-11.96% (5s segments) and 3.47%-9.62% (20s segments), respectively. To sum up, extracting the informative segments through the proposed activity and time ratios can effectively filter out the confounding segments, leading to the performance enhancement of factor-induced workload recognition can be enhanced.

D. Task Performance Evaluation from Eye-tracking Data

Eye-tracking data has been widely used to evaluate surgeons' skills [12], [35], which can help to improve the

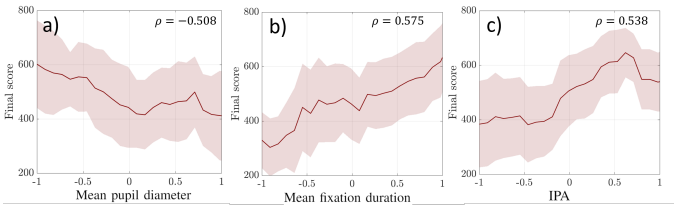


Fig. 8. Relationship between the simulator-defined *Final score* and three eye-tracking features extracted from trial data. The solid line indicates the mean value within a sliding window of size 0.1, and the shadow area represents the values within mean±std. PCC is shown at the top right corner. a) Mean fixation duration; b) Mean pupil diameter; c) IPA.

training procedure and overall performance. This paper also investigated various eye-tracking features for O^3 teleoperation performance evaluation. Different from the experiments in [12], [35] focusing on the comparison between expert and novice groups, all the operators were novices in teleoperating Canadarm2 robot on the O^3 simulator in this study.

1) Correlation with simulator-defined final score: First, we examined the correlation between the eye-tracking features and the simulator-defined *Final score* as introduced in Section III-B. To quantitatively evaluate the correlations, the Pearson correlation coefficient (PCC) was calculated by $\rho = \mathrm{cov}(x, y)/\sqrt{\mathrm{var}(x)\mathrm{var}(y)} \in [-1, 1]$, where cov and var represent covariance and variance, respectively.

Fig. 8 demonstrates three eye-tracking features that correlate with the *Final score* ($|\rho| \geq 0.5$). For eye movement related features as shown in Figs. 8(a), it can be found that superior performance on O^3 tasks was achieved when operators pay more attention, showing large *Mean fixation duration* during the task. For pupillary response related features, we calculated the average values of those features from Eye#0 and Eye#1. As illustrated in Figs. 8(b)&(c), *Mean pupil diameter* showed a negative correlation ($\rho = -0.505$) with *Final score*, while *IPA* representing cognitive workload had a positive correlation ($\rho = 0.538$). Similar to the findings in [6] that best performance was by adding moderated temporal stress, the performance measure degraded with a further increase of cognitive workload indicated by *IPA*, which reveals that adding the appropriate workload during space telerobotic training can help improve the training performance and facilitate the training process.

2) Subtask performance v.s. ratio of PoG on cameras: As operators' preference on cameras during grasping and docking are different, we explored the relationship between operators' interaction with visual displays (i.e., the ratio of PoG on different cameras within each trial) and their subtask performance (i.e., *Grasp score* and *Dock score*).

For grasping the research module, Fig. 9(a) shows that operators observed cameras #1&#3 located at the end-effector and the elbow link of Canadarm2 more often. On the other hand, cameras #3&#4 were used frequently during docking (see Fig. 9(c)). The correlations between the ratio of PoG and the simulator-defined measures are demonstrated in Fig. 9(b)(d). For cameras #1&#4, when the ratio of PoG was increased, operators' performance was dropped, which may reflect that operators were struggling in these subtasks. Conversely, we

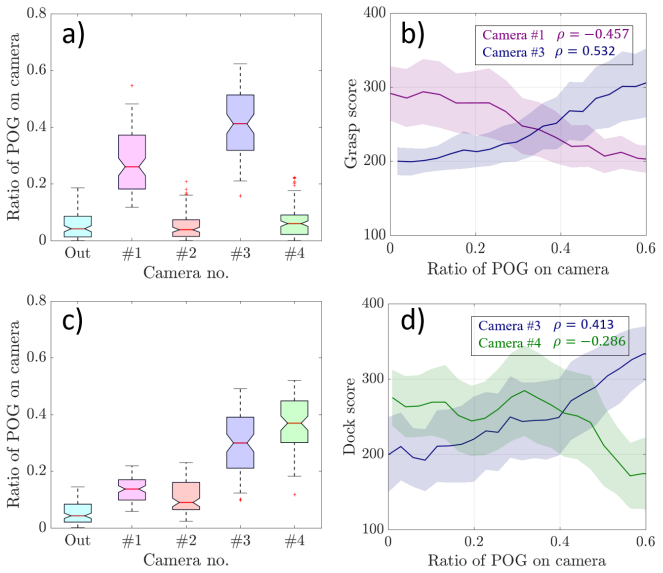


Fig. 9. Relationship between the ratio of PoG on cameras and performance (*Grasp score* and *Dock score*). The solid line in (b)(d) indicates the moving average with a sliding window of size 0.1, and the shadow area represents the values within mean±std. a) Camera ratio in grasping; b) Camera ratio v.s. grasp score; c) Camera ratio in docking; d) Camera ratio v.s. dock score.

can observe that there existed positive correlations between the ratio of PoG on camera#3 and {*Grasp score* ($\rho = 0.532$), *Dock score* ($\rho = 0.413$)}, which emphasizes that operators preferred operating Canadarm2 from its end-effector frame. More interestingly, camera#2 (on the wrist of Canadarm2) was rarely used by operators to accomplish the task, which we suspect is related to the poor lighting conditions for this camera's positioning. The above observations can help us to improve the user interface design in the future version.

VI. Discussions and Conclusions

This paper provided an in-depth investigation of factor-induced mental workload estimation and performance evaluation in space telerobotic training solely from eye-tracking data. Studies were performed on our O^3 training simulator, in which the operators' mental workload can be modulated by adding two confounding factors: time-pressure and latency. We first conducted the statistical analysis to examine how eye-tracking features change with respect to the existence of these factors, and then performed factor-induced workload identification using different eye-tracking features extracted from both the segment and trial data. Specifically, the workload identification using segment data enables us to offer opportunities for providing real-time feedback during the teleoperation training or real tasks.

With added time-pressure, we can observe the significant increases in {*Max fixation duration* and *IPA*} and decreases in {*Max saccade duration*, *Blink frequency*, and *Pupil deviation*}, which reflects that the operators' mental efforts were effectively modulated. In terms of the existence of latency, decreased *Mean fixation duration* and increased *Saccade frequency* were observed, and no significant differences were found in pupillary response features. This could indicate that

operators tend to change their control policy while their mental workload remained at a similar level. To improve the factor-induced workload recognition performance using segment data, we tried to disentangle it with task-specific workload by introducing the activity ratio and time ratio to distill the segments containing more information related to external confounding factors. Experimental results showed that our proposed small activity and time ratio (SATR) configuration consistently improved the results in both two-class and multi-class recognition tasks. Finally, eye-tracking based performance evaluation was conducted by exploring the correlation between eye-tracking features and simulator-defined performance measures. Results demonstrated that {*Mean fixation duration*, *Max saccade duration*, *Mean pupil diameter*, and *IPA*} correlated well with the final score. Additionally, the relationship between the ratio of PoG on each of the cameras and grasp/dock scores was examined, which can help to improve the user interface design.

The proposed eye-tracking based mental workload estimation and performance evaluation technology opens new possibilities for those needing to perform complex tasks with stress (e.g., pilots, surgeons, and drivers) and those populations with mild cognitive impairment. Accordingly, personalized training and assistive programs can be developed to effectively modulate their workload. For instance, instead of a fixed training plan, operators coping well with latency can be trained more with respect to time-pressure or other confounding factors. Besides, through the detection of anomaly workload, the safe operation can be further guaranteed. Although the scoring method is a reliable way for evaluating the performance in a training simulator, it is still paramount to estimate the performance in real tasks. Thus, the investigation of eye-tracking technology enables the transfer of teleoperation performance evaluation from simulated environments to real scenarios.

A limitation of this study is that only ten participants were involved in the experiments. Due to the individual difference, their performance, eye-tracking data, and the impact of time-pressure and latency on their mental workload will inevitably be affected. In the future, we will conduct more experiments on various time limits and latencies, which allow us to compare the differences across factors. Besides, future work will also focus on the fusion of multi-modal sensory data (e.g., body temperature, heart rate, EEG) for more accurate workload estimation and performance evaluation, thus providing real-time feedback related to operators' mental workload and various external factors experienced by operators.

REFERENCES

[1] T. Fong, J. Rochlis Zumbado, N. Currie, *et al.*, "Space telerobotics: unique challenges to human–robot collaboration in space," *Reviews of Human Factors and Ergonomics*, vol. 9, no. 1, pp. 6–56, 2013.

[2] J. C. Liou and N. L. Johnson, "Risks in space from orbiting debris," *Science*, vol. 311, pp. 340–341, 2006.

[3] J. Guo, Y. Liu, X. Kong, *et al.*, "Analysis of key cognitive factors in space teleoperation task," in *International Conference on Human-Computer Interaction*. Springer, 2019, pp. 249–258.

[4] S. Shao, Q. Zhou, and Z. Liu, "Study of mental workload imposed by different tasks based on teleoperation," *Int. J. Occup. Saf. Ergon.*, vol. 0, no. 0, pp. 1–11, 2020.

[5] E. Yang and M. C. Dorneich, "The emotional, cognitive, physiological, and performance effects of variable time delay in robotic teleoperation," *Int. J. Soc. Robot.*, vol. 9, no. 4, pp. 491–508, 2017.

[6] H. Singh, H. N. Modi, S. Ranjan, *et al.*, "Robotic surgery improves technical performance and enhances prefrontal activation during high temporal demand," *Ann. Biomed. Eng.*, vol. 46, no. 10, pp. 1621–1636, 2018.

[7] T. B. Sheridan, "Space teleoperation through time delay: Review and prognosis," *IEEE Trans. Rob. Autom.*, vol. 9, no. 5, pp. 592–606, 1993.

[8] D. Freer, Y. Guo, F. Deligianni, and G.-Z. Yang, "On-orbit operations simulator for workload measurement during telerobotic training," *arXiv preprint:2002.10594*, 2020.

[9] M. S. Prewett, K. N. Saboe, R. C. Johnson, *et al.*, "Workload in human-robot interaction: a review of manipulations and outcomes," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 53, no. 18. SAGE, 2009, pp. 1393–1397.

[10] G. Ziv, "Gaze behavior and visual attention: A review of eye-tracking studies in aviation," *Int. J. Aviat. Psychol.*, vol. 26, no. 3-4, pp. 75–104, 2016.

[11] S. Benedetto, M. Pedrotti, L. Minin, *et al.*, "Driver workload and eye blink duration," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 14, no. 3, pp. 199–208, 2011.

[12] T. Tien, P. H. Pucher, M. H. Sodergren, *et al.*, "Eye tracking for skills assessment and training: a systematic review," *J. Surg. Res.*, vol. 191, no. 1, pp. 169–178, 2014.

[13] G. N. Dimitrakopoulos, I. Kakkos, Z. Dai, *et al.*, "Task-independent mental workload classification based upon common multiband eeg cortical connectivity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1940–1949, 2017.

[14] J. Zhang, Z. Yin, and R. Wang, "Nonlinear dynamic classification of momentary mental workload using physiological features and narx-model-based least-squares support vector machines," *IEEE Trans. Hum. Mach. Syst.*, vol. 47, no. 4, pp. 536–549, 2017.

[15] C. Wu, J. Cha, J. Sulek, *et al.*, "Eye-tracking metrics predict perceived workload in robotic surgical skills training," *Hum. Factors*, vol. 62, no. 8, pp. 1365–1386, 2020.

[16] R. M. Hecht, A. B. Hillel, A. Telpaz, *et al.*, "Information constrained control analysis of eye gaze distribution under workload," *IEEE Trans. Hum. Mach. Syst.*, vol. 49, no. 6, pp. 474–484, 2019.

[17] D. B. Kaber, J. M. Riley, R. Zhou, and J. Draper, "Effects of visual interface design, and control mode and latency on performance, telepresence and workload in a teleoperation task," in *Proc. Hum. Factors Ergon. Soc. Annu. Mee.*, vol. 44, no. 5. SAGE, 2000, pp. 503–506.

[18] A. Khasawneh, H. Rogers, J. Bertrand, *et al.*, "Human adaptation to latency in teleoperated multi-robot human-agent search and rescue teams," *Automation in Construction*, vol. 99, pp. 265–277, 2019.

[19] S. Lu, M. Y. Zhang, T. Ersal, and X. J. Yang, "Workload management in teleoperation of unmanned ground vehicles: Effects of a delay compensation aid on human operators' workload and teleoperation performance," *Int. J. Hum. Comput. Interact.*, vol. 35, no. 19, pp. 1820–1830, 2019.

[20] H. N. Modi, H. Singh, F. Orihuela-Espina, *et al.*, "Temporal stress in the operating room: brain engagement promotes "coping" and disengagement prompts "choking"," *Ann. Surg.*, vol. 267, no. 4, pp. 683–691, 2018.

[21] E. Rendon-Velez, P. Van Leeuwen, R. Happee, *et al.*, "The effects of time pressure on driver performance and physiological activity: a driving simulator study," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 41, pp. 150–169, 2016.

[22] K. F. Van Orden, W. Limbert, S. Makeig, and T.-P. Jung, "Eye activity correlates of workload during a visuospatial memory task," *Hum. Factors*, vol. 43, no. 1, pp. 111–121, 2001.

[23] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Trans. Hum. Mach. Syst.*, vol. 48, no. 5, pp. 434–451, 2018.

[24] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," *Science*, vol. 154, no. 3756, pp. 1583–1585, 1966.

[25] W. Wang, Z. Li, Y. Wang, and F. Chen, "Indexing cognitive workload based on pupillary response under luminance and emotional changes," in *Proc. Int. Conf. Intelligent User Interfaces*, 2013, pp. 247–256.

[26] S. P. Marshall, "The index of cognitive activity: Measuring cognitive workload," in *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants*. IEEE, 2002, pp. 7–7.

[27] A. T. Duchowski, K. Krejtz, I. Krejtz, *et al.*, "The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation," in *Proc. SIGCHI Conf. Hum. Factor Comput. Syst.*, 2018, pp. 1–13.

[28] E. Kowler, "Eye movements: The past 25 years," *Vision Res.*, vol. 51, no. 13, pp. 1457–1483, 2011.

[29] C. M. Schulz, E. Schneider, L. Fritz, *et al.*, "Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment," *Br. J. Anaesth.*, vol. 106, no. 1, pp. 44–50, 2011.

[30] F. Volden, V. D. A. Edirisinghe, and K.-I. Fostervold, "Human gaze-parameters as an indicator of mental workload," in *Congress of the International Ergonomics Association*. Springer, 2018, pp. 209–215.

[31] J. Fan, J. W. Wade, A. P. Key, *et al.*, "Eeg-based affect and workload recognition in a virtual driving environment for asd intervention," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 1, pp. 43–51, 2017.

[32] J. Zhang, Z. Yin, and R. Wang, "Recognition of mental workload levels under complex human–machine collaboration by using physiological features and adaptive support vector machines," *IEEE Trans. Hum. Mach. Syst.*, vol. 45, no. 2, pp. 200–214, 2014.

[33] J. Currie, R. R. Bond, P. McCullagh, *et al.*, "Eye tracking the visual attention of nurses interpreting simulated vital signs scenarios: mining metrics to discriminate between performance level," *IEEE Trans. Hum. Mach. Syst.*, vol. 48, no. 2, pp. 113–124, 2017.

[34] L. Richstone, M. J. Schwartz, C. Seideman, *et al.*, "Eye metrics as an objective assessment of surgical skill," *Ann. Surg.*, vol. 252, no. 1, pp. 177–182, 2010.

[35] T. Tien, P. H. Pucher, M. H. Sodergren, *et al.*, "Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair," *Surg. Endosc.*, vol. 29, no. 2, pp. 405–413, 2015.

[36] H. Hamada, M. Inagami, T. Suzuki, *et al.*, "Effect of mental workload and aging on driver distraction based on the involuntary eye movement," in *Advances in Human Aspects of Transportation*. Springer, 2017, pp. 349–359.

[37] D. Su, Y. F. Li, and Y. Guo, "Precise gaze estimation for mobile gaze trackers based on hybrid two-view geometry," in *Proc. IEEE ROBIO*. IEEE, 2017, pp. 302–307.

[38] T. Kosch, M. Hassib, P. W. Woźniak, *et al.*, "Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload," in *Proc. SIGCHI Conf. Hum. Factor Comput. Syst.*, 2018, pp. 1–13.

[39] B. Pfleging, D. K. Fekety, A. Schmidt, and A. L. Kun, "A model relating pupil diameter to mental workload and lighting conditions," in *Proc. SIGCHI Conf. Hum. Factor Comput. Syst.*, 2016, pp. 5776–5788.

Yao Guo received his B.S. degree in automation and M.S. degree in communication and information system from Sun Yat-sen University, Guangzhou, China in 2011 and 2014, respectively. He earned his Ph.D. degree in robotic vision from the City University of Hong Kong, Hong Kong in 2018. His postdoctoral training was at the Hamlyn Centre for Robotic Surgery, Imperial College London, London, UK from 2018 to 2020.

Since 2020, he has been the tenure-track Assistant Professor with the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China. His main research interests include robotic vision, gait analysis, rehabilitation and assistive robotics, human-machine interaction, and machine learning algorithms in healthcare applications. He received the Best Conference Paper Award at the IEEE International Conference on Mechatronics and Automation (ICMA) 2016.

Daniel Freer is a researcher in the field of assistive robotics. He received his PhD in Robotics from Imperial College London in 2021, after a BEng degree from University of Pittsburgh in Bioengineering. Daniel is interested in a wide range of topics related to robotics and assistance, which range from Brain-Computer Interfaces (BCI) to autonomous vision-based robotic grasping.

Fani Deligianni received the M.Eng. (equivalent) degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2000, the M.Sc. degree in advanced computing from Imperial College London, London, U.K., in 2002, the Ph.D. degree in medical image computing from Imperial College London, in 2006, and the M.Sc. degree in neuroscience from University College London, London, in 2010.

She is currently a Lecturer with the School of Computing Science, University of Glasgow, Glasgow, U.K. Her interests include medical image/neuroimage computing, statistical machine learning and bioinformatics, and human motion analysis with wearable sensors and brain-computer interfaces.

Guang-Zhong Yang (Fellow, IEEE) was the Director and the Co-Founder of the Hamlyn Centre for Robotic Surgery and the Deputy Chairman of the Institute of Global Health Innovation, Imperial College London, London, U.K., where he also holds a number of key academic positions, such as the Director and the Founder of the Royal Society/Wolfson Medical Image Computing Laboratory, the Co-Founder of the Wolfson Surgical Technology Laboratory, and the Chairman of the Centre for Pervasive Sensing. He is currently the Founding Dean of the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China. His main research interests are in medical imaging, sensing, and robotics. In imaging, he is credited for a number of novel MR phase contrast velocity imaging and computational modeling techniques that have transformed in vivo blood flow quantification and visualization. These include the development of locally focused imaging combined with real-time navigator echoes for resolving a respiratory motion for high-resolution coronary angiography, as well as the MR dynamic flow pressure mapping for which he received the ISMRM I. I Rabi Award. He pioneered the concept of perceptual docking for robotic control, which represents a paradigm shift of learning and knowledge acquisition of motor and perceptual/cognitive behavior for robotics, as well as the field of the body sensor network (BSN) for providing personalized wireless monitoring platforms that are pervasive, intelligent, and context-aware.

Dr. Yang is a fellow of the Royal Academy of Engineering, the Institution of Engineering and Technology (IET), and the American Institute for Medical and Biological Engineering (AIMBE). He was a recipient of the Royal Society Research Merit Award. He is listed in The Times Eureka "Top 100" in British Science. Professor Yang is the founding editor of Science Robotics (http://robotics.sciencemag.org/) – a journal of the Science family dedicated to the latest advances in robotics and how it enables or underpins new scientific discoveries. He was awarded a CBE in the Queen's 2017 New Year Honour for his contribution to biomedical engineering.