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

I. Index of Pupillary Activity and the Detection of PoG on four cameras

The pseudo-code of the calculation of Index of Pupillary Activity (IPA) [1] is summarized in Algorithm 1.

Algorithm 1: Index of Pupillary Activity (IPA) [1]

Input: Pupil diameter sequence x_d and duration t_n seconds;
Output: IPA score for x_d;

(1) Wavelet decomposition: $\texttt{cD}_2$ = `wavedec`(x_d,`level`=2,'`sym16`'), where `cD` represents the detailed coefficients;
(2) Normalization: Normalize $\texttt{cD}_2$ by $1/2^k$, where k is the level ($k = 2$);
(3) Modulus maxima detection: $\texttt{cD}_{2m}$ =`modmax`($\texttt{cD}_2$);
(4) Calculate threshold λ from $\texttt{cD}_{2m}$: $\lambda = \sigma\sqrt{2\log n}$, where σ and n are the standard deviation and length of $\texttt{cD}_{2m}$;
(5) Sign function: $\texttt{cD}_{2t} = |\texttt{sign}\,(\texttt{cD}_{2t} - \lambda)|$;
(6) Compute IPA: $\texttt{IPA=numOfPos}(\texttt{cD}_{2t})/t_n$, where `numOfPos` returns the number of positive elements.

In order to determine the 2D point of gaze on four cameras in terms of the head movement, we extracted the boundaries of the visual displays in real-time using image-processing techniques. In Fig. S-A, the detected purple boundaries are highlighted, where the remaining parts are displayed in grayscale. By accurately localizing the cross boundaries, the 2D PoG on four camera displays can be determined.

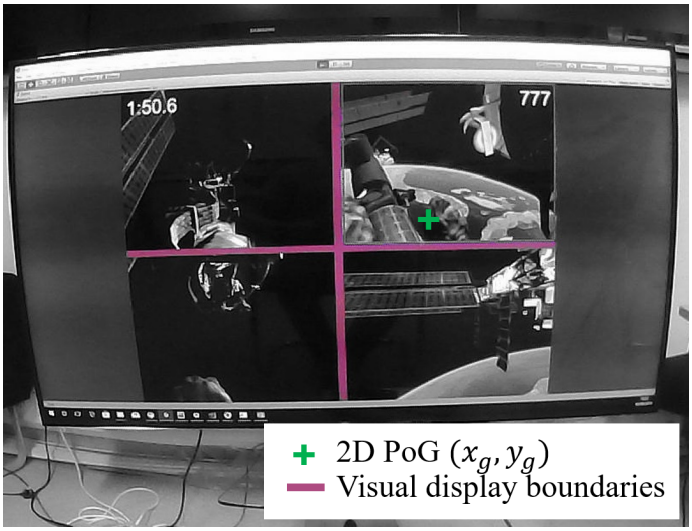


Fig. S-A. Determine PoG on four camera displays through the detection of purple boundaries on the user interface. The detected boundaries are highlighted with their original color, where other parts are shown in grayscale.

II. Statistical analysis of performance measures and eye-tracking features

The descriptions of the eye-tracking features and p-value results of different features on a single confounding factor are listed in Table S-A. Besides, we listed results of simulator-defined performance measures, which have been presented in our previous work [2]. It can be observed that, with added *Time pressure*, operators finished the O^3 subtasks more efficiently, where the *Grasp time* and *Dock time* were significantly decreased. With added *Latency*, operators collided with obstacles more often. For other simulator-defined performance measures, there was no statistical significance can be found.

TABLE S-A
P-value results of different eye-tracking features with and without a single confounding factor

Category	Eye-tracking features	Descriptions	Time pressure		Latency	
			p-value	$\frac{\mu_{w/}}{\mu_{w/o}}$	p-value	$\frac{\mu_{w/}}{\mu_{w/o}}$
Task measure	Grasp time	Elapsed time from the start of simulation to finish grasping	0.016*	↓	0.733	-
	Grasp angle error	Angle error between desired and actual grasping poses	0.902	-	0.943	-
	Grasp distance error	Euclidean distance between the desired and actual grasping positions	0.285	-	0.139	-
	Grasp score	Simulator-defined score for the grasping subtask	0.664	-	0.205	-
	Dock time	Elapsed time from the grasping to the finish of docking	0.000**	↓	0.157	-
	Dock angle error	Angle error between the desired and actual docking poses	0.386	-	0.817	-
	Dock distance error	Euclidean distance between the desired and actual docking positions	0.741	-	0.675	-
	Dock score	Simulator-defined score for the docking subtask	0.691	-	0.973	-
	Final score	Overall score per trial	0.290	-	0.175	-
	Collisions	Number of collisions with obstacles per trial/segment	0.285	-	0.016*	↑
Eye movement related features	Fixation frequency	Number of fixations per second	0.842	-	0.118	-
	Mean fixation duration	Mean time duration of the fixations per trial/segment	0.153	-	0.039*	↓
	Max fixation duration	Max time duration of the fixations per trial/segment	0.031*	↑	0.164	-
	Saccade frequency	Number of saccades per second	0.872	-	0.025*	↑
	Mean saccade duration	Mean time duration of saccades per trial/segment	0.188	-	0.331	-
	Max saccade duration	Max time duration of the saccades per trial/segment	0.042*	↓	0.680	-
	Mean saccade speed	Mean speed of eye saccades per trial/segment	0.463	-	0.335	-
	Max saccade speed	Max speed of eye saccades per trial/segment	0.518	-	0.424	-
Blink related features	Blink frequency	Number of blinks per second	0.028*	↓	0.318	-
	Mean blink duration	Mean time duration of the blinks per trial/segment	0.697	-	0.548	-
	Max blink duration	Max time duration of the blinks per trial/segment	0.568	-	0.994	-
Pupillary response related features (Eye#0 - Eye#1)	IPA#0	Index of Pupillary Activity of eye#0	0.009**	↑	0.789	-
	Mean pupil#0 diameter	Mean pupil#0 diameter per trial/segment	0.255	-	0.548	-
	Pupil#0 deviation	(Max - Min) pupil#0 diameter per trial/segment	0.047*	↓	0.552	-
	Max pupil #0 diameter	Max pupil#0 diameter per trial/segment	0.268	-	0.917	-
	IPA#1	Index of Pupillary Activity of eye#1	0.049*	↑	0.673	-
	Mean pupil#1 diameter	Mean pupil#1 diameter per trial/segment	0.474	-	0.909	-
	Pupil#1 deviation	(Max - Min) pupil #1 diameter per trial/segment	0.039*	↓	0.549	-
	Max pupil #1 diameter	Max pupil #1 diameter per trial/segment	0.556	-	0.854	-

• $^{**}p \leq .005$, $^{*}0.005 < p \leq .05$; ↑, ↓ indicate the changes of mean value $\mu_{w/}$ compared to $\mu_{w/o}$; - represents no significant difference.

III. Factor-induced Workload Identification from Eye-tracking Data

Detailed comparison results in discriminating the workload induced by adding a single factor with different segments $\{2s, 5s, 10s, 20s\}$ and the trial data under LOSO protocol are reported in Table S-B. 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 = positive) under LOSO protocol respectively. Then we calculated the final `F1` score to evaluate the performance.

Table S-C lists the recognition results on more challenging four-level recognition task under LOSO protocol, where the comparison results were performed by using the segment data of size $\{2s, 5s, 10s, 20s\}$ and the trial data, respectively. The average 5-fold validation accuracy (`ValAcc`) during training and the test accuracy (`TestAcc`) under LOSO evaluation protocol were calculated. The detailed comparison of multi-class recognition with the original segment data and the segments with five proposed configures (**SAR**, **LAR**, **STR**, **LTR**, **SATR**) is demonstrated in Fig. S-B. Figs. S-C(a)-(e) shows the confusion matrices for the results (`TestAcc`) using *ANOVA* features extracted from original $20s$ segments (33.85%), the trial data (49.32%), and the proposed configurations {**SAR** (40.02%), **STR** (39.97%), **SATR** (43.68%)}.

TABLE S-B
TWO-CLASS RECOGNITION RESULTS ON SINGLE FACTOR UNDER LOSO PROTOCOL

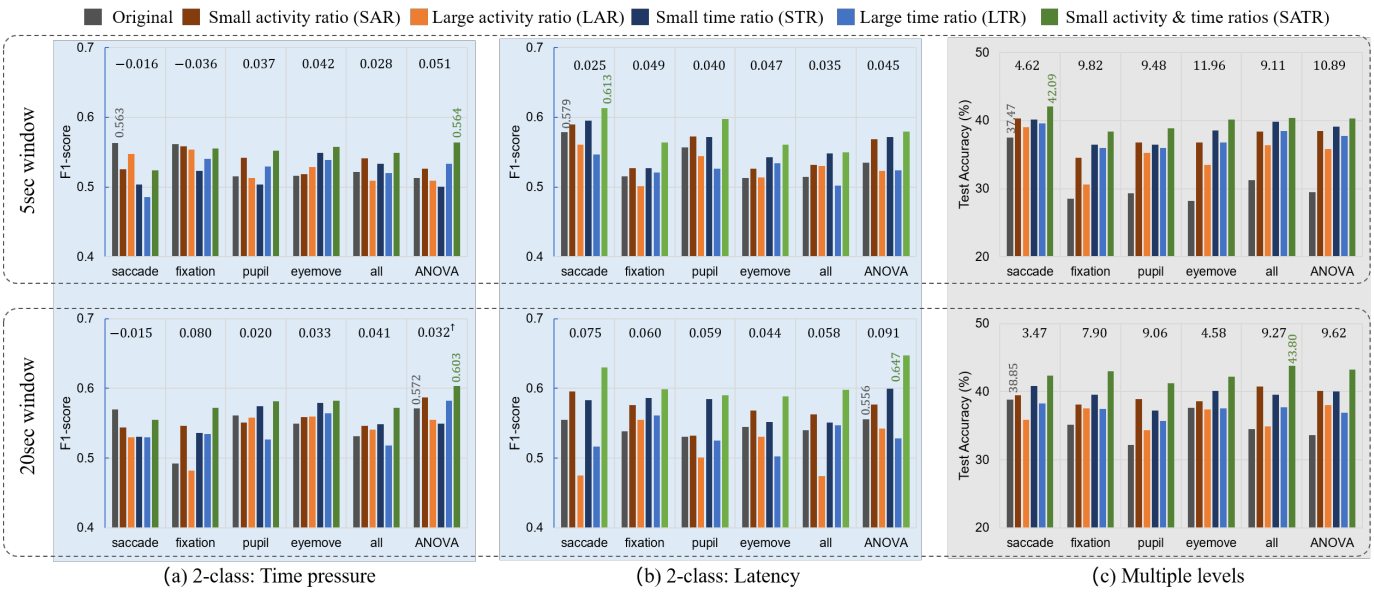
		Segment data												Trial data		
		$2s$			$5s$			$10s$			$20s$					
Factor	Features	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
Time Pressure (tp)	blink	.378	.498	.429	.484	.499	.492	.521	.512	.517	**.525**	**.529**	**.527**	.503	.477	.488
	saccade	.539	.499	.519	**.619**	.516	.563	.536	.510	.523	.613	**.532**	**.570**	.515	.552	.532
	fixation	.523	.505	.514	**.576**	**.548**	**.562**	.531	.515	.526	.483	.501	.492	.622	.651	.636
	pupil	.523	.495	.509	.526	.504	.515	.524	.523	.523	**.574**	**.549**	**.561**	.604	.566	.584
	eyemove	.464	.456	.460	.525	.508	.516	.534	**.535**	.534	**.575**	.527	**.550**	.631	.638	.634
	all	.505	.498	.502	**.547**	.498	.521	.521	**.523**	.522	.545	.518	**.532**	.689	.701	.695
	ANOVA	.491	.481	.486	.522	.505	.513	.535	.517	.526	**.604**	**.542**	**.572**	**.749**	**.751**	**.750**
Factor	Features	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
Latency (lat)	blink	.437	**.501**	.467	.444	.466	.455	.418	.453	.435	**.477**	.483	**.480**	.544	.560	.552
	saccade	.575	.534	.554	**.584**	**.574**	**.579**	.559	.549	.554	.563	.548	.555	.545	.546	.545
	fixation	**.550**	**.547**	**.548**	.514	.515	.515	.507	.505	.506	.538	.538	.538	.559	.548	.553
	pupil	.539	.537	.538	**.560**	**.554**	**.557**	.509	.511	.510	.533	.529	.531	.628	.626	.627
	eyemove	.543	.542	.543	.516	.511	.513	.488	.483	.485	**.545**	**.544**	**.544**	.606	.562	.583
	all	.502	.502	.502	.514	.516	.515	.509	.505	.505	**.539**	**.540**	**.540**	.611	.569	.589
	ANOVA	.517	.518	.517	.537	.534	.535	.529	.524	.526	**.557**	**.554**	**.556**	**.707**	**.708**	**.707**

• **Bold** indicates the highest value in each row (feature), 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, respectively.

TABLE S-C
RECOGNITION ACCURACIES ON FOUR LEVELS CLASSIFICATION PARADIGM

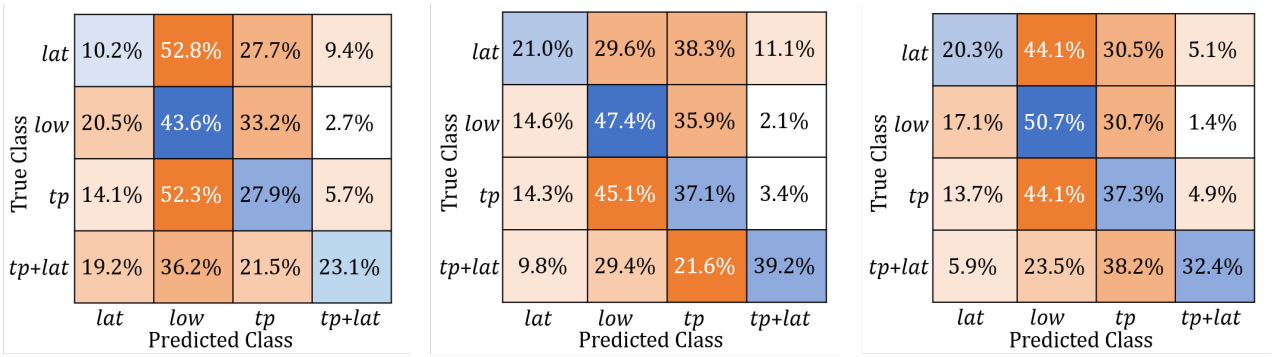
Features	Segment data								Trial data	
	$2s$		$5s$		$10s$		$20s$			
	ValAcc	TestAcc	ValAcc	TestAcc	ValAcc	TestAcc	ValAcc	TestAcc	ValAcc	TestAcc
blink	**48.61**	**38.49**	46.29	34.73	42.86	33.42	42.19	35.89	32.09	33.14
saccade	**50.43**	37.92	48.73	37.47	43.24	35.89	42.59	**38.85**	35.63	34.37
fixation	**60.43**	34.79	55.57	28.52	47.59	32.48	49.62	**35.10**	37.56	37.03
pupil	**71.88**	30.86	65.51	29.36	56.76	30.12	50.59	**32.17**	32.29	36.31
eyemove	**72.13**	30.58	58.85	28.19	55.27	30.26	47.17	**37.61**	36.04	38.21
all	**76.64**	30.49	65.01	31.25	62.58	28.45	52.81	**34.53**	34.52	39.16
ANOVA	**81.67**	29.05	67.20	29.45	63.54	30.67	54.70	**33.58**	**48.37**	**49.32**

• Each element indicates the recognition accuracy (%); **Bold** indicates the highest value in each row (feature) among segment data of different sizes, and Underline is the maximum in each column . Blue background highlights the best result for recognizing each factor from segment data and trial data.

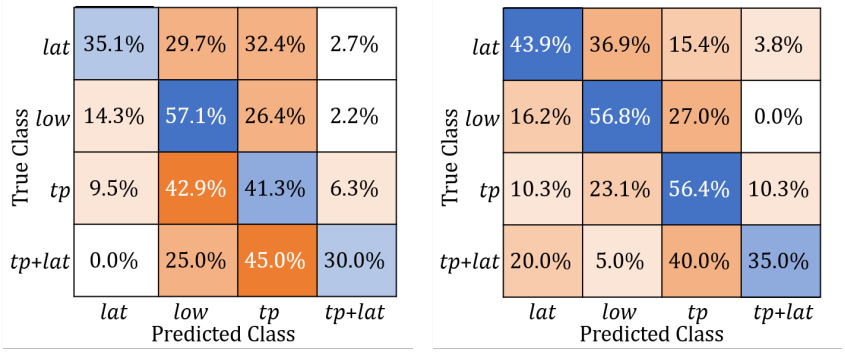


(a) 2-class: Time pressure (b) 2-class: Latency (c) Multiple levels

Fig. S-B. Comparison 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.



(a) Original segment data (b) Small activity ratio (c) Small time ratio



(d) Small activity and time ratio (e) Trial data

Fig. S-C. Confusion matrices for multi-class workload recognition using the *ANOVA* features extracted from a) original 20*s* segment data; b) 20*s* segments with small activity ratio; c) 20*s* segments with small time ratio; d) 20*s* segments with small activity and time ratio; and e) trial data.

REFERENCES

[1] 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.

[2] 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.