EPiC Series in Computing

Volume 102, 2024, Pages 126–139

Proceedings of Conference on Cognitive and Computational Aspects of Situation Management 2023

Uncertainty-Aware Visual Workload Estimation for Human-Robot Teams

Joshua Bhagat Smith*[1], Simone Angelo Toribio*[2], Julie A. Adams[1]

[1] Oregon State University, Corvallis, Oregon, USA
bhagatsj@oregonstate.edu, adamsjuli@oregonstate.edu
[2] University of Minnesota, Minneapolis, Minnesota, USA
torib005@umn.edu

Abstract

Human-robot teams operate in uncertain environments and need to accomplish a wide range of tasks. A dynamic understanding of the human's workload can enable fluid interactions between team members. A system that seeks to adapt interactions for a human-robot team needs to quantify the distribution of workload across the different workload components. A workload assessment algorithm capable of estimating the demand placed on the human's visual resources is required. Further, adaptive systems will benefit from measures of uncertainty, as these measures inform interaction adaptations. Two machine learning methods' capacity to estimate visual workload for a human-robot team operating in a non-sedentary supervisory environment are analyzed. A key finding is that the uncertainty-aware method outperforms the other approach.

1 Introduction

Human-robot teams operating in dynamic environments will require robots to accurately estimate the human's internal state (e.g., workload level). These estimates must account for real-world complexities to enable fluid interactions between the robot and its human teammate. Workload is defined as a ratio of internal resources required to complete a task and the available resources dedicated to that task [29]. Workload can be decomposed into different components (i.e., cognitive, visual, speech, auditory, gross motor, fine motor, and tactile) based on the type of resources a task utilizes [24]. Incorporating accurate workload models enhances the robot's understanding of its human teammate, allowing the robot to manage the human's workload by intelligently selecting appropriate interaction modalities (e.g., visual vs. auditory).

Effective adaptation will only be achieved when a robot can quantify the extent to which each workload component contributes to an individual's overall workload. Prior work developed a multi-dimensional workload algorithm that trains separate machine learning methods for each workload component [13]. This approach is theoretically capable of assessing the specific

*These authors contributed equally.

demand each workload component places on the human's overall workload, but only incorporates cognitive, auditory, speech, gross motor, fine motor, and tactile workload [13] [7] [2]. The absence of a visual workload method is a core deficiency that leaves the algorithm uninformed.

Future human-robot teams will operate in dynamic environments where each team member must be able to move freely; thus, physiological sensors required for human workload estimation must be mobile and worn by the human. Wearable eye trackers (e.g., Pupil Core [18]) allow individuals to move freely in the environment while leaving their vision relatively untouched. Prior work demonstrated that ocular metrics (e.g., pupil diameter, blink rate, fixation duration) correlate with changes in cognitive workload (e.g., [12] [22] [10]), but these metrics also share a natural connection to visual workload. Distinguishing between the contributions of cognitive workload and visual workload is critical when adapting interactions with the human.

Consider the wildland firefighting domain. Fireline construction incurs a high visual workload because the human must visually identify where to dig, but requires little cognitive processing. Comparatively, firefighters monitoring sensor feeds (e.g., thermal imagery) from a unmanned aerial vehicle surveying the forest requires significant amounts of both cognitive and visual resources. Enumerating these differences helps the robot determine how best to communicate with the human. Further, uncertainty contextualizes component-wise estimates as the components rely on overlapping metrics.

Incorporating a visual workload method into the multi-dimensional workload algorithm enables the algorithm to fully diagnose the type of workload a human experiences. This manuscript compares two machine learning methods for estimating visual workload. Both methods were trained using data from a human supervising a remotely-located aerial robot. Robot interactions were simulated through a modified NASA Multi-Attribute Task Battery-II (MATB-II), which incorporated tasks of varying difficulties to induce different workload levels. Ground truth values were derived using a human performance modeling tool, which provided anchors for each task's workload and produced continuous workload values. Predicting a continuous workload value helps determine the human's proximity to a high or low workload state.

2 Background

Humans have a limited capacity for processing information and making decisions. This capacity can be analyzed by viewing how humans work from a resource management perspective. Humans exhibit individual differences [29] that vary day-to-day and are dependent upon a range of external factors, such as expertise, training, and fatigue.

2.1 Workload Overview

Workload metrics can take the form of (1) subjective questionnaires, (2) performance-based measures, and (3) objective methods derived from physiological signals. The most widely used subjective questionnaire is the NASA Task Load Index (NASA TLX) [11], which is typically administered upon task completion. Prior work examined relationships between physiological metrics using NASA-TLX; however, these methods are susceptible to human subjectivity [6]. Another commonly chosen workload measure is performance-based metrics (e.g., [31] [21]), which tend to be task specific, confounded by competency levels, and often do not generalize.

Machine learning can learn the relationship between physiological metrics and an individual's underlying workload (e.g., [2], [10], [30]). Interestingly, these objective workload methods often use subjective questionnaires and performance-based measures as the ground truth. Prior work either focused on estimating overall workload for visual tasks [10], or detecting workload levels

for simple cognitive workload for visual tasks [30]. Further, many of these methods only perform discrete workload classification (e.g., low vs. high) workload [30].

IMPRINT Pro is a human performance modeling tool that allows users to construct complex task networks, where tasks can be organized sequentially, concurrently, and hierarchically [24]. Each task's workload value is derived using pre-defined component-specific anchor values. The resulting human performance model generates continuous workload values for each workload component that can be used as the ground truth in a machine learning process [13]. Utilizing continuous workload values enables a workload estimation method to both classify workload into relative workload conditions (i.e., high vs. low) and inform the robot about the human's proximity to a high or low workload.

High and low workload values have different implications on a human's internal state. A high workload value, *overload* (OL), occurs when a task requires a large amount of resources and an individual only has a small amount of resources available [29]. OL can decrease performance because the human has insufficient resources to perform a given task. A low workload value, *underload* (UL), occurs when a task requires relatively few resources and the individual has a large amount of available resources. UL can present a problem when humans become unengaged in their work because that can lead to reduced alertness and lowered situational awareness [29]. Thresholds for OL and UL vary between individuals and tasks, and nominal workload values between these thresholds are referred to as *normal load* (NL).

2.2 Visual Workload

Many objective workload methods rely on ocular metrics [10] [30]; however, none specifically estimate visual workload. Visual workload is the amount of work induced by a task that can include registering visual stimuli, reading, or visually scanning an area. Ocular metrics share a natural connection to visual workload; however, many of these metrics also correspond to cognitive workload and have only been evaluated in that context (e.g., [10], [31], [21]). Pupil dilation, blink frequency, fixation duration, and saccades are among popular metrics used for estimating cognitive workload (e.g., [10] [20]). These metrics are sensitive and diagnostic for cognitive workload [22], but their utility for visual workload estimation has not been evaluated.

Pupil dilation is the most commonly used ocular metric; however, ambient lighting conditions heavily influence the diameter of an individual's pupil, and this metric's usefulness significantly diminishes with high lighting variations. Recent techniques mitigate this negative impact by using discrete wavelet transformations [5], which produced two metrics: the Index of Cognitive Activity and the Index of Pupillary Activity (IPA). The Index of Cognitive Activity requires a proprietary algorithm [23]. IPA is an open-source equivalent to the Index of Cognitive Activity [5]. Both metrics are as sensitive to changes in workload and are more robust to ambient lighting changes. Further evidence is required to fully evaluate these metrics' utility for visual workload estimation.

Gaze Entropy is the Shannon Entropy of fixations [26]. Prior research demonstrated that pilots display non-deterministic visual scanning patterns when their aircraft is in an error-free state, and deterministic scanning patterns during emergencies [4]. Gaze entropy is sensitive to changes in cognitive workload, but has not been evaluated in the context of visual workload [30].

Many of these metrics have been successfully used for visual task recognition in sedentary environments (e.g., [19], [27]). Visual task recognition and workload estimation are distinct problems, but the success of these task recognition algorithms demonstrates the potential of ocular metrics for visual workload estimation. However, these ocular metrics used in these algorithms required the human's head to be still and required a high-frequency data stream [17].

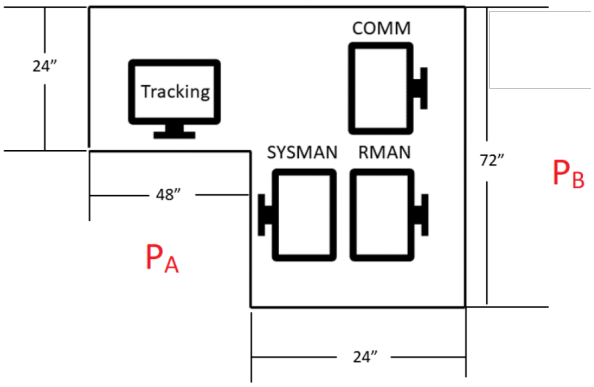


Figure 1: The Modified NASA MATB-II physical layout. PA and PB represent the areas participants walked between. SYSMAN: System Monitor station. RMAN: Resource Management station. COMM: Communications station.

Recent efforts use deep learning to produce more accurate gaze features in non-stationary environments by correcting for situational factors using the forehead's angular velocity [8].

3 Methods

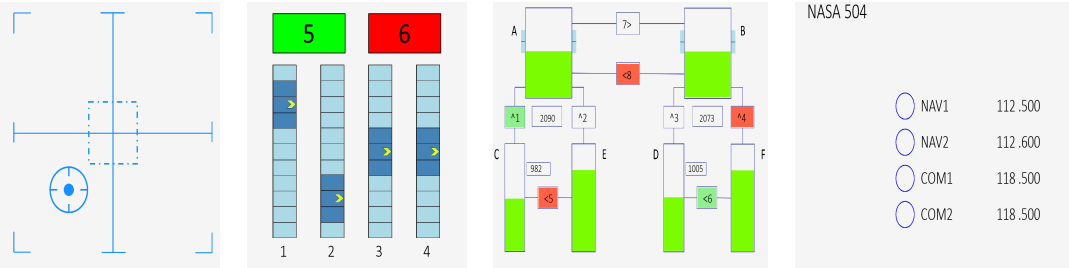
A mixed-subjects supervisory user evaluation was designed to manipulate participants' workload based on task density [28]. This evaluation was designed to evaluate standard machine learning methods' ability to accurately estimate workload for all workload components. Two machine learning methods were evaluated: 1) a neural network, and 2) a Gaussian process. Neural networks learn complex and highly non-linear interactions between multiple features and have been effective at estimating other workload components [13] [2]; thus, hypothesis H_1 predicts that the neural network will have better performance. Nevertheless, prior work suggests that the presented ocular metrics strongly correlate with changes in workload for visually demanding tasks [10]; thus, hypothesis H_2 predicts that both methods will be capable of distinguishing between relative workload levels.

3.1 Experimental Design

Table 1: Independent variables.

Type	Variable
within-subjects	Tasks Task-density (i.e., workload)
between-subjects	Workload ordering

This evaluation manipulated tasks, task density, and workload ordering as independent variables, see Table 1. The task environment was the NASA Multi-Attribute Task Battery-II (MATB-II) [3], which simulates a supervisory-based human-robot team. Participants completed a single 52.5-minute trial, which consisted of seven consecutive 7.5 minute intervals. Task density manipulated the number of tasks initiated during a specific interval [28]. Workload was



(a) Tracking (b) System Monitoring (c) Resource Management (d) Communication

Figure 2: The NASA MATB-II Tasks.

elicited by increasing or decreasing the NASA MATB-II tasks' frequency in three levels, each corresponding to a relative workload level (i.e., UL, NL, and OL). Three task density orderings were used to manipulate workload, and ensured that each workload transition (e.g., UL-NL, OL-UL) occurred exactly once:

- O1: UL-NL-OL-UL-OL-NL-UL
- O2: NL-OL-UL-OL-NL-UL-NL
- O3: OL-UL-OL-NL-UL-NL-OL

The supervisory task environment consisted of a modified version of the NASA MATB-II [3], which required a human operator to supervise a simulated remotely-piloted aircraft. The NASA MATB-II consists of four tasks: tracking, system management, resource management, and communication. The communication task was split into two separate tasks, communication and communication response, in order to model the communication itself and any verbal response. The original NASA MATB-II required participants to remain stationary at a single workstation. The NASA MATB-II was modified to physically separate each task to require participants to walk between two stations, as depicted in Figure 1. Each NASA MATB-II task had a separate dedicated monitor, positioned such that the participant was unable to visually see more than two monitors simultaneously. This visual hindrance ensured that participants walked around the environment to complete the overall mission. The required equipment (e.g., joystick or keyboard) to complete each task was placed in front of the respective computer monitor.

The tracking task, depicted in Figure 2a, required participants keep the circle with a blue dot in the middle of the cross-hairs using a joystick. The underload condition required a single 45s session of manual tracking, with automated tracking consuming the remaining time. The overload condition had two 12s manual tracking sessions every minute, while the normal load condition had one 20s session every minute.

The system monitoring task, shown in Figure 2b, required monitoring two colored lights and four gauges. If the L5 light turned green or the L6 light turned red, the value was out of range and required resetting. The four gauges had a randomly moving up and down indicator that typically remained in the middle. Participants reset a gauge if it was out of range (i.e., too high or too low). These items were reset by pressing the corresponding number key on the keyboard's top row. The underload condition had only one out of range instance in the entire 7.5 minute session, overload had fifteen instances per minute, and normal load had five instances per minute.

The resource management task included six fuel tanks (A-F) and eight fuel pumps (1-8), shown in Figure 2c. The arrow by the fuel pump's number indicated the direction fuel flows. Participants were to maintain the fuel levels of Tanks A and B by turning the fuel pumps on or off. Fuel Tanks C and D had finite fuel levels, while Tanks E and F had an infinite supply. A pump turned red when it failed. The underload condition had 2 minutes of manual resource management with zero pumps failing, while the remaining time was automated. The normal load condition required manual resource management during the last 3.5 minutes, with at most two pumps failing every minute. The overload condition required manual resource management for the entire 7.5 minute condition, with two or more pumps failing every minute.

The communications task, depicted in Figure 2d, required listening to air-traffic control requests for radio changes. These requests were similar to: "NASA 504, please change your COM1 radio to frequency 127.550." The original MATB communications task required no speech, but a required verbal response was added. An example response is: "This is NASA 504 tuning my COM 1 radio to frequency 127.550." Participants changed the appropriate radio to the specified frequency by selecting the radio and using arrows to change its frequency. Communications not directed to the participant's aircraft, indicated by the call sign, were to be ignored. The underload condition contained a single request with one response task, the overload contained three requests with at least two response tasks every minute, while normal load contained up to two requests, with one response task per minute.

Finally, the Pupil Core eye tracker [18] was connected to the desktop computer with a 10-foot USB cable, allowing participants to move freely between stations. The participants were required to walk around the tables to the other station (e.g., from PA to PB shown in Figure 1) whenever a ping sound occurred. The participants were free to move between the tasks at any time, but the ping sound enforced a mandatory transition to the other workstation. The underload condition contained two walk requests, the overload condition incorporated seven requests per minute, and normal load had two requests per minute. Task timings and occurrences were chosen such that the correct workload condition, or task density, was elicited.

3.2 Dependent Variables

The human performance modeling tool IMPRINT Pro was used to model tasks for each workload level and ordering prior to conducting the evaluation. Tasks were assigned a workload value using IMPRINT Pro's anchors [24] and modelled overtime to produce continuous workload values. These values served as the ground truth for all machine learning methods.

The visual workload estimation method incorporates metrics collected using the Pupil Labs Core eye tracker [18]. The Core eye tracker streams pupil metric data at 200 HZ, but streams gaze metric data at 30 HZ. Pupil Lab's eye tracking software derives other metrics (e.g., blinks, fixations, saccades) from these two data streams. A total of twenty four metrics were used, where each metric corresponds to one of four metric types: *pupil*, *blink*, *fixation*, and *saccade*.

Eight pupil metrics were incorporated into the visual workload method: the mean, standard deviation (std.), and maximum pupil diameter, as well as the mean IPA [5]. Pupil metrics for each eye were considered separately.

Two blink metrics were incorporated: blink rate, and blink latency. Blink rate is defined as the number of blinks per minute and blink latency is defined as the time between consecutive blinks. Prior work established that blink rate increases with visual workload and decreases with cognitive workload [22]; thus, this metric can inform the machine learning method of the resources being utilized (i.e., cognitive or visual).

Fixations are defined by the points where an individual's gaze is stationary, and saccades are

defined by the jumps, or transitions, between these fixation points [22]. The minimum duration for determining a fixation varies between applications. The Pupil Labs software calculates fixations using a dispersion-based algorithm (i.e., degrees of visual angle), where fixations occur when the gaze location is stationary for a user-defined amount of time. The preset minimum threshold of 80 milliseconds and a maximum threshold of 300 milliseconds were chosen based on the eye tracker's documentation [18]. Five fixation metrics were incorporated: fixation frequency, gaze entropy [26], as well as the mean, std., and maximum fixation duration. Seven saccade metrics were incorporated into the method: saccade frequency, the mean, std., and maximum of both saccade duration, as well as saccade speed [22].

Two metrics were chosen to evaluate the machine learning methods: Root Mean Squared Error (RMSE) and Spearman correlation. RMSE represents the average distance between the estimated workload and the IMPRINT Pro model's workload. The Spearman correlation captures the degree to which the estimated workload changes when the IMPRINT Pro workload changes, even when the RMSE is large. It is important to contextualize the presented correlation values. Prior work identified correlation values between ± 0.7 and ± 0.9 as "High positive (negative) correlation", values between ± .50 and ± .70 as "Moderate positive (negative) correlation", values between ± .30 and ±.50 as "Low positive (negative) correlation", and values between .00 to ± .30 as having a "negligible correlation" [15].

In situ workload ratings required a response on a Likert scale (1 - very low, 5 - very high) for each workload component. In situ workload ratings were verbally administered six minutes into the trial and every 7.5 minutes after the initial rating. These subjective ratings serve as a secondary workload measure and were used to verify the machine learning methods' results.

3.3 Visual Workload Method

A prior visual task recognition algorithm achieved its highest accuracy levels using a 30s window size [1]; thus, all ocular metrics were processed using a 30s sliding window. Data within the window was standardized by subtracting the mean (μ) and dividing by the std. (σ): $x_{normalized} = (x - \mu)/\sigma$. Normalized ocular metrics served as the machine learning methods' input features and the IMPRINT Pro model served as the ground truth workload values.

Two machine learning methods were evaluated: (1) a shallow neural network, and (2) a Gaussian process. Neural networks can classify overall workload for visually demanding tasks [21] [31], and provide granular estimates for other workload components [13] [7]. A neural network with 5 layers was implemented using PyTorch, where each hidden layer had 128 nodes. The network was trained using the Adam optimizer with a batch size of 64. Early stopping was employed to prevent overfitting. Likewise, Gaussian processes have successfully produced workload estimates for cognitive workload (e.g., [21], [31]). The Gaussian process was implemented using GPyTorch [9]. Exact Gaussian Processes are notoriously difficult to train, as they are intractable to solve for large datasets; thus, a stochastic variational Gaussian process (SVGP) was implemented [14]. Both machine learning methods were validated using leave-one-subject-out cross validation.

3.4 Procedure

The participants completed a consent form and a demographic questionnaire upon arrival, after which participants were fitted with several wearable sensors, including the Pupil Core eye tracker [18]. A tutorial video described how to accomplish the tasks, which was followed by a 10-minute training session. The training session cycled through the five tasks, with each task occurring for a 1-minute period, and repeated the cycle one additional time. The 52.5-minute

Table 2: RMSE and Spearman Correlation for the leave-one-subject-out cross validation for Visual Workload. Note that the term "All" refers to results for all orderings combined. The best results is bolded.

Ordering	Method	RMSE	Spearman Correlation			
		Mean (Std)	Mean (Std)	Min	Median	Max
All	**Neural Net**	4.391 (0.461)	0.471 (0.171)	0.050	0.505	0.702
	SVGP	**4.108 (0.388)**	**0.486 (0.200)**	-0.055	**0.537**	0.762
O1	**Neural Net**	4.385 (0.438)	0.478 (0.204)	0.050	0.531	0.702
	SVGP	**4.117 (0.389)**	**0.507 (0.203)**	0.008	**0.589**	0.762
O2	**Neural Net**	4.328 (0.527)	**0.485 (0.171)**	0.187	0.541	0.696
	SVGP	**4.018 (0.434)**	0.480 (0.239)	-0.055	**0.571**	0.730
O3	**Neural Net**	4.467 (0.396)	0.445 (0.109)	0.245	0.492	0.630
	SVGP	**4.195 (0.301)**	**0.464 (0.135)**	0.109	**0.493**	0.644

trial switch the tasks rapidly and sometimes overlapped tasks in order to emulate real-world scenarios. The participants completed a post-session questionnaire upon finishing the trial.

Sixty-four participants (37 male, 24 female, and 3 non-binary) completed the experiment. The mean age and std. were 29.80 and 10.24. Thirty-four held a high school degree, fourteen held an undergraduate degree, fourteen held a master's degree, and five held a doctorate degree. Participants indicated the number of hours they use a desktop or laptop per week, as computer experience may impact task performance. The majority of participants (45) indicated that they use computers for more than eight hours per week. Participants rated their video game skill level on average as 4.75 (std. = 2.62) on a Likert scale (1-little to 9-expert). The results did not exhibit meaningful differences across these factors.

4 Results

The Friedman analysis of variance by ranks test was used to analyze the results when there are more than two groups. If significant differences exist, the Wilcoxon signed-rank test was applied. These non-parametric statistical tests ensure that the outcomes were unaffected by the error distribution across participants. The RMSE and Spearman Correlation values for both machine learning methods are provided in Table 2.

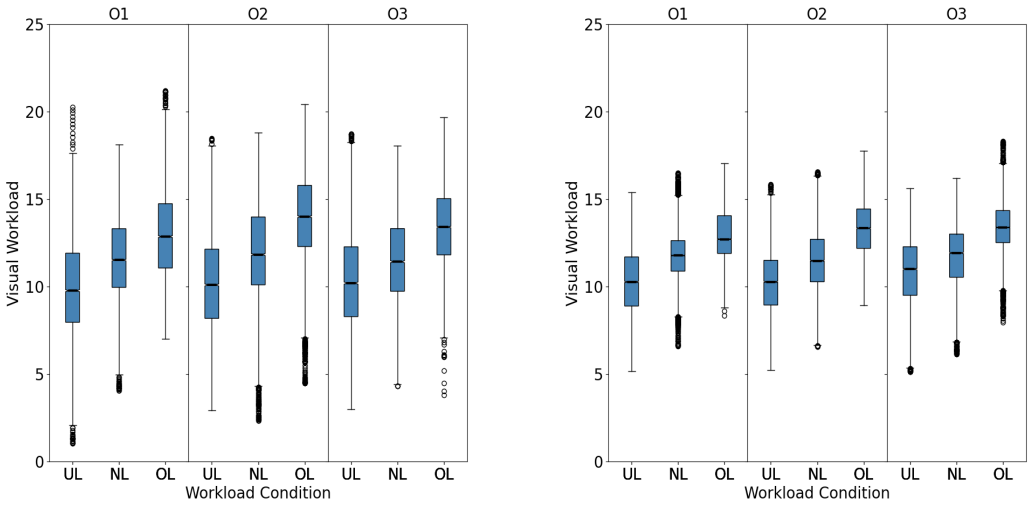
The SVGP outperformed the neural network across the majority of independent variables. The SVGP achieved a lower RMSE across all (i.e., All) and each workload ordering when compared to the neural network. The Wilcoxon signed-rank test indicated that the SVGP's RMSE was significantly lower for All ($p < 0.01$), as well as for each ordering (O1 and O2: $p < 0.01$, O3: $p < 0.05$). Similar patterns existed with Spearman correlation, as the SVGP's correlation was significantly higher for All and O3 ($p < 0.05$). The SVGP's correlation value was higher for O1 and lower for O2, but were not statistically significant. The Cohen d effect size for the RMSE results was All: 0.654, O1: 0.622, O2: 0.612, and O3: 0.738, while small effect size was found for correlation All: 0.121, O1: 0.135, O2: 0.012, and O3: 0.369.

Neither machine learning method's performance was significantly impacted by workload ordering. A Friedman analyses showed that the SVGP's RMSE and Spearman correlation had no statistically significant differences across workload orderings, and similar results were found for the neural network.

A box plot of the estimated workload values produced by the neural network and SVGP

Table 3: Visual Workload In Situ Workload Ratings.

Ordering	Workload		
	UL	NL	OL
All	2.59 (1.009)	3.05 (1.089)	3.800 (1.038)
O1	2.439 (0.923)	3.550 (0.686)	3.952 (1.024)
O2	1.895 (1.048)	2.923 (1.109)	3.611 (1.036)
O3	2.238 (1.091)	2.923 (1.187)	3.81 (1.078)



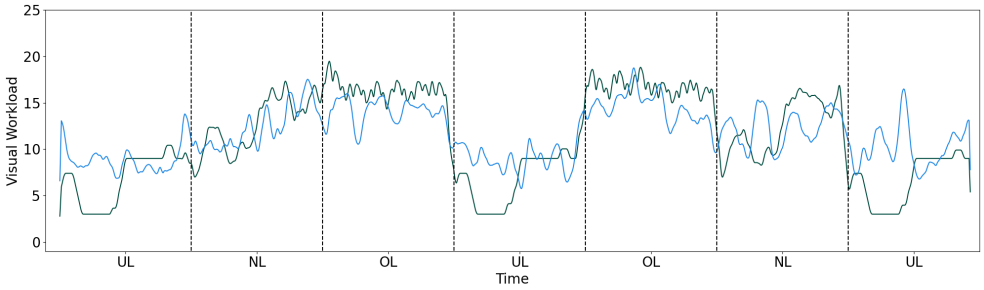
(a) Neural Network

(b) SVGP

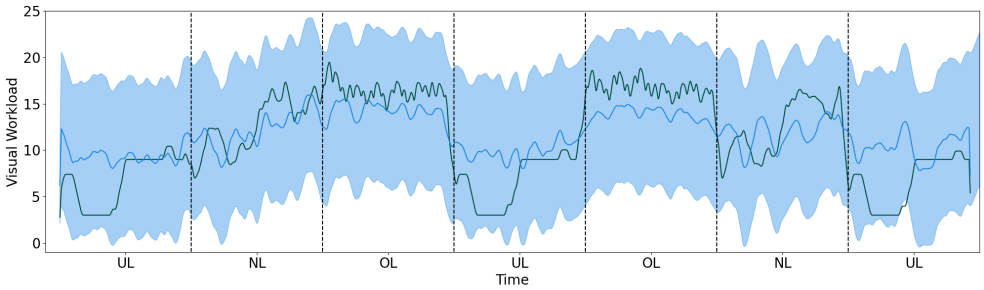
Figure 3: Quantile plots of the (a) neural network, and the (b) Gaussian Process by ordering for mean estimated visual workload, with the additional outliers (white circles).

for each workload level (i.e., UL, NL, OL), and each workload ordering is presented in Figure 3. The Friedman analysis showed that there was a statistically significant difference between the workload conditions ($p < 0.01$) for both methods across all three workload orderings. The Wilcoxon ranked tested verified the difference between each pair of workload conditions (i.e., UL-NL, NL-OL, and UL-OL) was statistically significant ($p < 0.01$) for both methods across all three workload orderings. Further, the Cohen d results showed the effect size of these differences were all large (i.e., $d > 0.5$). Additionally, these results mirror the participants' ratings, as demonstrated by their reported in situ workload ratings presented in Table 3. The Friedman analysis and pairwise Wilcoxon tests for the in situ workload ratings showed significant differences across all independent variables, with large effect sizes.

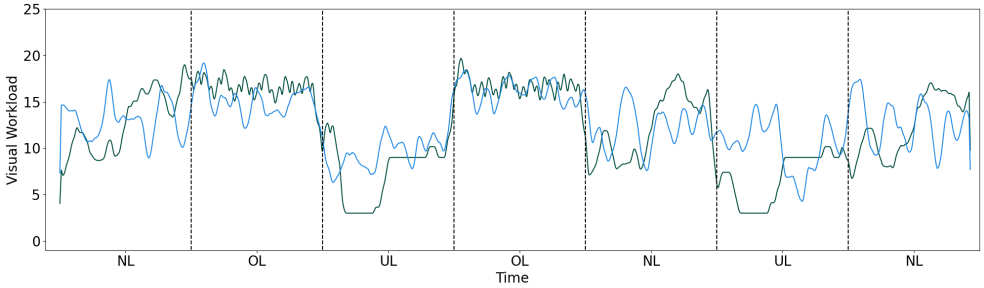
Visual workload estimates for two participants over time, from two different orderings, for the neural network (Figures 4a and c) and the SVGP (Figures 4b and d) are provided. Figures 4a and b represent the participant with the highest average correlation across both machine learning methods, and the participant who exhibited a median average correlation and slightly below average RMSE is presented in Figures 4c and d. The shaded regions represent one standard deviation from the mean estimated visual workload.



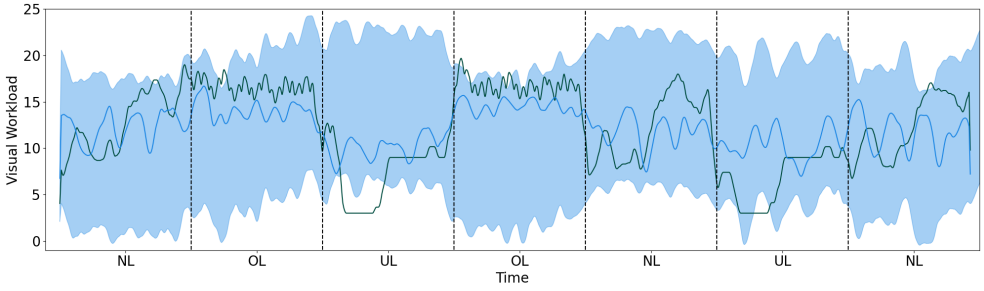
(a) Neural network results for the participant with the highest average correlation (RMSE: 3.48, Corr: 0.69).



(b) SVGP results for the participant with the highest average correlation (RMSE: 3.45, Corr: 0.76).



(c) Neural network results for a participant with median average correlation (RMSE: 3.92, Corr: 0.53).



(d) SVGP results for a participant with median average correlation (RMSE: 3.77, Corr: 0.56).

Figure 4: Visual workload estimates for the two example participant. (Green) Ground truth IMPRINT Pro values for visual workload. (Blue) A machine learning method's workload estimates.

These figures demonstrate that both machine learning methods accurately estimate visual workload, especially during the NL and OL conditions. However, the SVGP's uncertainty region is large for the duration of the trial, demonstrating that this method never produces highly confident (i.e., low uncertainty) visual workload estimates.

The participant in Figure 4a experienced three UL conditions during the trial. The neural network grossly overestimates the visual workload during the last UL condition (i.e., 45-minute mark). Similar over-estimations are made during both UL conditions (i.e., 20 and 37-minute mark) for the median-correlation participant (see Figure 4c). The methods generally detect the decrease in visual workload, but fail to capture the full magnitude of that decrease.

The performance difference between the highest correlation participant (i.e., 0.73) and the median correlation participant (i.e., 0.55) is difficult to visually discern. The median correlation participant has increasingly noisy workload estimates during the last two intervals when compared to the highest correlation participant. These noisy estimates are a likely source of performance difference between these two participants.

5 Discussion

Hypothesis H_1 was not supported, in fact the results strongly support the opposite conclusion since the SVGP achieved higher performance levels. The SVGP achieved a lower RMSE and higher Spearman correlation for all and individual workload orderings. Further, machine learning method choice had a large effect on RMSE.

Generally, the SVGP appears to be more capable at estimating visual workload. However, both machine learning models were able to detect the statistically significant differences between relative workload conditions; thus, the SVGP's higher performance did not translate into an increased ability to differentiate between workload conditions. These findings strongly support Hypothesis H_2. Further, the in situ workload results confirm that participants experienced significant differences in visual workload between conditions. These differences existed for all workload orderings, and for each individual ordering.

A key limitation of both machine learning methods is the inability to accurately estimate workload during the UL condition, which is likely the result of superfluous task monitoring during workload transitions. Verifying participants execute tasks in the same order as the IMPRINT Pro model, at the expected time, is non-trivial and highly uncertain. Task uncertainty can cause misalignment between the objective workload estimation and the IMPRINT Pro workload model, leading to poor performance during the training process. Misalignment is more likely to occur during transition periods where participants may expect tasks to continue with the same frequency; thus, participants may continue to engage in excessive monitoring and inflate their workload. Further, the in situ workload ratings capture the individual's workload for a single time period (e.g., 7.5-minute interval), but may not represent the workload experienced during these transition periods. It is difficult to ensure that participants do not engage in excessive task monitoring during transitions because enforcing hard constraints on participant behavior introduces bias and does not represent realistic operational conditions.

A key benefit to SVGPs is uncertainty quantification, because uncertainty can provide additional context for informing to be developed system adaptations. Uncertainty naturally corresponds to confidence. Adaptations informed by highly confident workload estimates are justifiable, because the robot estimate of the humans' internal state will be more accurate. Adaptations informed by low confidence workload estimates may be counter-productive when informed by potentially inaccurate estimates. A limitation of the SVGP is this high uncertainty, which will need to be resolved in the future.

Eye tracker slippage and variability of the participants eye physiology are likely sources of this large uncertainty region, as they both increase noise in the physiological metrics. Prior work demonstrated gaze estimation error increased significantly (i.e., 0.8-3.1 degrees) when the Pupil Core eye tracker slipped over time [25], and these noisy estimates are most likely caused by slippage in the eye tracker. Reliably gathering the ocular metrics is crucial to the method's performance. Differences in eye physiology, lash direction, baseline pupil size, as well as face shape and viewing angle all negatively impact the reliable data collection [16].

The multi-dimensional workload assessment algorithm requires a visual workload estimation method in order to assess the specific demand each workload component places on the human's overall workload. Both machine learning methods accurately estimated visual workload. The SVGP's primary benefit is uncertainty quantification, as highly confident workload estimates are required for the robot to take definitively helpful actions. The presented results suggest that more work is needed to ensure the SVGP's uncertainty is less impacted by individual differences and noise, such that the uncertainty is more useful for future robot adaptivity algorithms.

6 Conclusion

Two machine learning algorithms were developed using physiological metrics extracted from wearable sensors. Both methods were able to accurately estimate visual workload and differentiate between relative workload conditions, but the SVGP achieved a significantly lower RMSE and higher Spearman correlation. The future ability for a robot to adapt to the human's workload in a helpful manner requires differentiating which workload channels are overtaxed, which the SVGP method did. Future robot adaptivity algorithms can be further informed by robust uncertainty quantification, as highly confident workload estimates verify the robot's decisions are based on accurate information. Future work will investigate how to improve uncertainty quantification of the presented visual workload estimation method.

References

[1] P. Baskaran, J. Bhagat Smith, and J.A. Adams. Visual task recognition for human-robot teams. In *IEEE International Conference on Human-Machine Systems*, pages 1–6. IEEE, 2022.

[2] Joshua Bhagat Smith, Prakash Baskaran, and Julie A Adams. Decomposed physical workload estimation for human-robot teams. In *IEEE International Conference on Human-Machine Systems*. IEEE, 2022.

[3] J Raymond Comstock Jr and Ruth J Arnegard. The multi-attribute task battery for human operator workload and strategic behavior research. Technical report, NASA, 1992.

[4] Carolina Diaz-Piedra, Hector Rieiro, Alberto Cherino, Luis J Fuentes, Andres Catena, and Leandro L Di Stasi. The effects of flight complexity on gaze entropy: An experimental study with fighter pilots. *Applied ergonomics*, 77:92–99, 2019.

[5] Andrew T Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[6] Seth Elkin-Frankston, Bethany K Bracken, Scott Irvin, and Michael Jenkins. Are behavioral measures useful for detecting cognitive workload during human-computer interaction? In *Advances in The Human Side of Service Engineering*, pages 127–137. Springer, 2017.

[7] Julian Fortune, Jamison Heard, and Julie A Adams. Real-time speech workload estimation for intelligent human-machine systems. In *Human Factors and Ergonomics Society Annual Meeting*, volume 64, pages 334–338. Sage publications, 2020.

[8] Wolfgang Fuhl and Enkelejda Kasneci. A multimodal eye movement dataset and a multimodal eye movement segmentation analysis. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–7, 2021.

[9] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in Neural Information Processing Systems*, 31, 2018.

[10] Yao Guo, Daniel Freer, Fani Deligianni, and Guang-Zhong Yang. Eye-tracking for performance evaluation and workload estimation in space telerobotic training. *IEEE Transactions on Human-Machine Systems*, 52(1):1–11, 2021.

[11] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 904–908. Sage publications, 2006.

[12] Jamison Heard, Caroline E Harriott, and Julie A Adams. A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems*, 48(5):434–451, 2018.

[13] Jamison Heard, Rachel Heald, Caroline E Harriott, and Julie A Adams. A diagnostic human workload assessment algorithm for collaborative and supervisory human–robot teams. *IEEE Transactions on Human-Robot Interaction*, 8(2):1–30, 2019.

[14] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, page 282. Citeseer, 2013.

[15] Dennis E Hinkle, William Wiersma, and Stephen G Jurs. *Applied Statistics for the Behavioral Sciences*, volume 663. Houghton Mifflin college division, 2003.

[16] Kenneth Holmqvist, Saga Lee Örbom, Ignace TC Hooge, Diederick C Niehorster, Robert G Alexander, Richard Andersson, Jeroen S Benjamins, Pieter Blignaut, Anne-Marie Brouwer, Lewis L Chuang, et al. Eye tracking: empirical foundations for a minimal reporting guideline. *Behavior Research Methods*, 55(1):364–416, 2023.

[17] Mir Riyanul Islam, Shaibal Barua, Mobyen Uddin Ahmed, Shahina Begum, Pietro Aricò, Gianluca Borghini, and Gianluca Di Flumeri. A novel mutual information based feature set for drivers' mental workload evaluation using machine learning. *Brain Sciences*, 10(8):551, 2020.

[18] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1151–1160, 2014.

[19] Conor Kelton, Zijun Wei, Seoyoung Ahn, Aruna Balasubramanian, Samir R Das, Dimitris Samaras, and Gregory Zelinsky. Reading detection in real-time. In *Symposium on Eye Tracking Research & Applications*, pages 1–5, 2019.

[20] Krzysztof Krejtz, Justyna Żurawska, Andrew T Duchowski, and Szymon Wichary. Pupillary and microsaccadic responses to cognitive effort and emotional arousal during complex decision making. *Journal of Eye Movement Research*, 13(5), 2020.

[21] Pujitha Mannaru, Balakumar Balasingam, Krishna Pattipati, Ciara Sibley, and Joseph Coyne. Cognitive context detection in uas operators using eye-gaze patterns on computer screens. In *Next-Generation Analyst IV*, volume 9851, pages 133–143. SPIE, 2016.

[22] Gerhard Marquart, Christopher Cabrall, and Joost de Winter. Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, 3:2854–2861, 2015.

[23] Sandra P Marshall. The index of cognitive activity: Measuring cognitive workload. In *IEEE Conference on Human Factors and Power Plants*, pages 7–7. IEEE, 2002.

[24] Diane K Mitchell. Mental workload and ARL workload modeling tools. Technical report, Army Research Lab Aberdeen Proving Ground MD, 2000.

[25] D.C. Niehorster, T. Santini, R.S. Hessels, I. Hooge, E. Kasneci, and M. Nyström. The impact of

slippage on the data quality of head-worn eye trackers. *Behavior Research Methods*, 52:1140–1160, 2020.

[26] Brook Shiferaw, Luke Downey, and David Crewther. A review of gaze entropy as a measure of visual scanning efficiency. *Neuroscience & Biobehavioral Reviews*, 96:353–366, 2019.

[27] Namrata Srivastava, Joshua Newn, and Eduardo Velloso. Combining low and mid-level gaze features for desktop activity recognition. *Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–27, 2018.

[28] Mathew B Weinger, Oliver W Herndon, Mark H Zornow, Martin P Paulus, David M Gaba, and Larry T Dallen. An objective methodology for task analysis and workload assessment in anesthesia providers. *Anesthesiology*, 80(1):77–92, 1994.

[29] Christopher D Wickens, Sallie E Gordon, Yili Liu, and J Lee. *An Introduction to Human Factors Engineering*, volume 2. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[30] Chuhao Wu, Jackie Cha, Jay Sulek, Tian Zhou, Chandru P Sundaram, Juan Wachs, and Denny Yu. Eye-tracking metrics predict perceived workload in robotic surgical skills training. *Human factors*, 62(8):1365–1386, 2020.

[31] Yutaka Yoshida, Hayato Ohwada, Fumio Mizoguchi, and Hirotoshi Iwasaki. Classifying cognitive load and driving situation with machine learning. *International Journal of Machine Learning and Computing*, 4(3):210, 2014.