














HuBar: A Visual Analytics Tool to Explore Human Behavior based on fNIRS in AR Guidance Systems

Sonia Castelo , Joao Rulff , Parikshit Solunke , Erin McGowan , Guande Wu , Iran Roman , Roque Lopez , Bea Steers , Qi Sun , Juan Bello , Bradley Feest , Michael Middleton , Ryan Mckendrick, and Claudio Silva 

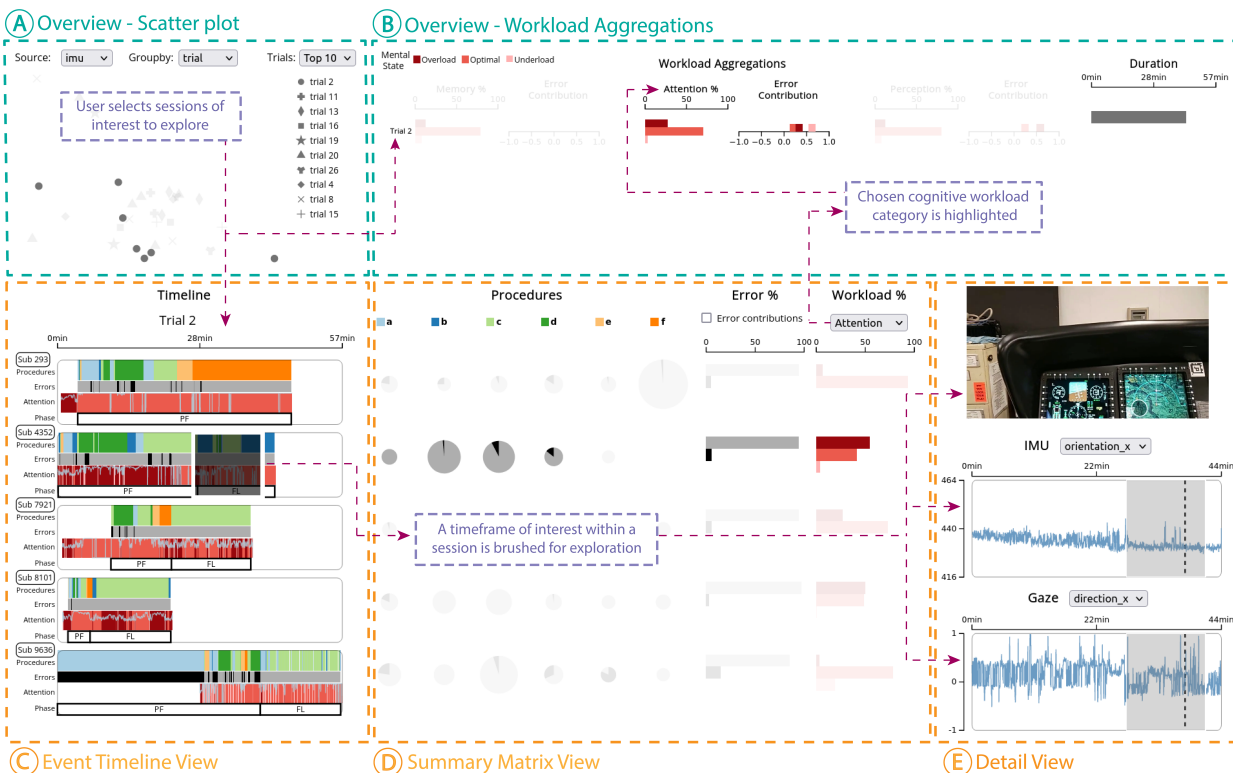


Fig. 1: *HuBar* is a visual analytics system that offers a hierarchical set of visualizations designed to analyze performer behavior in augmented reality (AR) assistance tasks by enabling multi-perspective analysis of multimodal time-series data. The Overview includes the Scatter Plot View (A), which offers three different projections of performer sessions, revealing multidimensional clusters and patterns while enabling filtering and selection of sessions. The Workload Aggregation View (B) summarizes performers' cognitive workloads, session durations, and workload-error correlations for the chosen groups. The Event Timeline View (C) aligns multiple time series (procedures, mental workload, errors, task phases) collected during performer sessions along a shared time axis, enabling comparison across sessions and exploration by brushing to update linked views. The Summary Matrix View (D) facilitates analysis of procedure frequency, error proportion, overall errors, and mental state distribution within and across sessions for selected workload categories. The Detail View (E) enables in-depth exploration of individual sessions with synchronized video and time series visualizations, supporting brushing for seamless navigation and analysis.

Abstract—The concept of an intelligent augmented reality (AR) assistant has significant, wide-ranging applications, with potential uses in medicine, military, and mechanics domains. Such an assistant must be able to perceive the environment and actions, reason about the environment state in relation to a given task, and seamlessly interact with the task performer. These interactions typically involve an AR headset equipped with sensors which capture video, audio, and haptic feedback. Previous works have sought to facilitate the development of intelligent AR assistants by visualizing these sensor data streams in conjunction with the assistant's perception and reasoning model outputs. However, existing visual analytics systems do not focus on user modeling or include biometric data, and are only capable of visualizing a single task session for a single performer at a time. Moreover, they typically assume a task involves linear progression from one step to the next. We propose a visual analytics system that allows users to compare performance during multiple task sessions, focusing on non-linear tasks where different step sequences can lead to success. In particular, we design visualizations for understanding user behavior through functional near-infrared spectroscopy (fNIRS) data as a proxy for perception, attention, and memory as well as corresponding motion data (acceleration, angular velocity, and gaze). We distill these insights into embedding representations that allow users to easily select groups of sessions with similar behaviors. We provide two case studies that demonstrate how to use these visualizations to gain insights about task performance using data collected during helicopter copilot training tasks. Finally, we evaluate our approach by conducting an in-depth examination of a think-aloud experiment with five domain experts.

Index Terms—Perception & Cognition, Application Motivated Visualization, Temporal Data, Image and Video Data, Mobile, AR/VR/Immersive, Specialized Input/Display Hardware.

1 INTRODUCTION

The concept of an AI-assisted task guidance system, which guides a user through a task using wearable sensors to detect objects and actions, is quickly shifting from science fiction to an impending reality. The potential applications of task guidance systems include physical tasks across a wide variety of domains such as medicine, mechanics, and military endeavors. Such a system could introduce tasks to trainees starting a new role and track their performance improvement over time, both for their own benefit and for the retrospective analysis of their peers. It could also serve as a second pair of eyes for domain experts, increasing task efficiency, especially during repetitive or stressful tasks.

In recent years, enormous advancements in machine perception and reasoning, along with hardware innovations, have made it possible to begin developing robust AI-assisted task guidance systems [28, 30]. This is a complex undertaking, requiring several heterogeneous sensors and machine learning models to work together to perceive the physical environment and reason about object state changes relevant to a given task. These systems typically involve an augmented reality (AR) headset, which superimposes graphics onto the performer's real-world environment and collects data relevant to their behavior, (e.g. egocentric video, audio, gaze, hand interactions) [7, 19]. Moreover, these data can be augmented with external sensors that gather information about human behavior, such as sensors to perform functional near-infrared spectroscopy (fNIRS), a popular technique for studying brain activity which is widely used to quantify mental workload [4, 34]. This performer behavior and mental workload data enable task guidance systems to adapt task instructions based on the performer's mental state (for clarity, we refer to subjects using the AR system to perform tasks during a session as "performers" and subjects using *HuBar* to analyze data as "*HuBar* users").

The recent development of increasingly sophisticated AR headsets (e.g., Microsoft HoloLens, Meta Quest, Apple Vision Pro) provides the hardware necessary for AI-assisted task guidance, and has also piqued the interest of stakeholders who could benefit from such a system. This increase in popularity has also prompted initiatives to collect task performance data from subjects with varying expertise levels. However, turning these data into useful insights requires intuitive systems which enable developers and researchers to understand human behavior at scale and under heterogeneous constraints.

Previous efforts have proposed approaches to explore performer actions (e.g. position and gaze) over time using custom visualizations [8, 10]. These approaches, however, lack mechanisms to understand the performer's mental state and how it correlates to their actions. Furthermore, these previous works do not explore comparison between individuals with different levels of expertise at the given task. More detailed performer modeling could make AR systems more adaptable and aide in coaching or performance report applications, especially if this performer modeling is situated in the context of data describing the surrounding environment.

Challenges in modeling performer behavior. To effectively model performer behavior, we must determine a method of summarizing and comparing performer behavior across sessions. This necessitates a meaningful way to compare multimodal time series data (e.g. gaze origin and direction, acceleration, angular velocity, fNIRS sensor readings) of different durations. This is a nontrivial task, especially since two performers may both successfully complete the same task by performing the same steps in different orders, or even by repeating some

steps. Moreover, performer behavior modeling requires a robust method for visualizing any correlation between cognitive workload (e.g. from fNIRS sensor data) and the sensor data streams capturing the motion of the performer.

Our Approach. We propose *HuBar*, a visual analytics tool for summarizing and comparing task performance sessions in AR based on performer behavior and cognitive workload using fNIRS, gaze, and inertial measurement unit (IMU) data. The *HuBar* interface is composed of a hierarchy of four visual components that allow the *HuBar* user to compare recorded task guidance sessions at varying levels of detail. At the overview level, *HuBar* users can compare sessions based on IMU, gaze, or fNIRS data, explore aggregated metrics for performer perception, attention, and memory workload, and select sessions of interest (see Sec. 4.2). *HuBar* users can then use the Event Timeline View to understand correspondences between task procedures, human errors, workload effects, and task phases for selected sessions (see Sec. 4.2). The Summary Matrix View increases the level of granularity of this analysis by showcasing how human error varies with each task procedure. Finally, the Detail View shows the video, IMU, and gaze data for selected portions of a given session (see Sec. 4.2). All views are linked and interactive. In short, our tool facilitates post-hoc analysis of task guidance in AR through visualizations that highlight similarities and differences in performer behavior between multiple task sessions, flag human errors in task performance, and display how the performer's cognitive workload level responds to events in the physical environment.

Our design was inspired by requirements and intermittent feedback from developers of AR systems and experts that create and evaluate these systems in the context of the Defense Advanced Research Projects Agency's (DARPA) Perceptually-enabled Task Guidance (PTG) program [12]. To summarize, **our main contributions are:**

- An interactive visualization tool, *HuBar*, containing a hierarchy of visualizations that facilitate the exploration and comparison of performer behavior at varying levels of detail, specifically highlighting the correlations between cognitive workload, IMU, gaze, and actions during task performance. This interface was designed to enable the comparison of multimodal time series data corresponding to interleaved task procedures of differing durations.
- We illustrate the value of *HuBar* through two case studies that demonstrate how domain experts leverage the tool as an after-action report and in a coaching scenario using real-world data.
- We validate our design decisions through interviews with 5 domain experts with extensive experience (collectively) in human factors, fNIRS, biovisualization, neuroinformatics design, and AR.

This paper is organized as follows: Sec. 2 reviews the relevant literature on human motion analysis based on time series, measuring workload effects in AR environments, and human behavior based on fNIRS data. Sec. 3 describes the data. Sec. 4 specifies the requirements we aim to achieve and describes *HuBar* in detail, including each aspect of the visualization design. Sec. 5 outlines two case studies in which *HuBar* proves useful to experts in our chosen domain, followed by an expert interview and discussion of the feedback we received on our system. Sec. 6 includes a discussion of limitations of our system, potential future works, and concluding remarks.

2 RELATED WORK

2.1 Human Behavior Analysis based on Time Series

The analysis of human behavior using time series data from various sensors, including wearable and AR devices, is well-studied. Activity recognition, a core application, leverages data from IMUs found in smartphones, watches, and earbuds to estimate and predict body movements over time, illustrating the potential of wearable sensors in capturing dynamic human data [27].

Key to analyzing human behavior is the extraction of meaningful features from sensor data. Studies have demonstrated the use of advanced

Authors are with the New York University (NYU) and Northrop Grumman Corporation (NGC). E-mails: {s.castelo, jlruflff, pss442, erin.mcgowan, guandewu, irr2020, rlopez, bs3639, qs2053, jpbello, csilva}@nyu.edu, {bradley.feest, michael.middleton, ryan.mckendrick}@ngc.com

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxx

techniques, such as time series shapelets, to segment behavior activities from sensor data [16, 21, 32]. Fulcher's work further underscores the significance of integrating multiple data streams for a holistic view of behavioral patterns [14].

2.2 Visualization Tools for Human Behavior Analysis based on Time Series

Various visualization tools have been introduced to analyze and interpret sensor data for human behavior analysis. Chan et al.'s Motion Browser for analyzing upper limb movements [11] and Xu et al.'s ensemble of techniques for multimodal data analysis [38] represent significant advancements. These tools facilitate understanding of muscle coordination, behavior distribution, and interdependence among behavioral variables through sophisticated visual analytics. Notably, a study by Öney et al. provided insight into best practices for visualizing time series data collected by an AR headset using gaze data [44]. This system utilized both qualitative and quantitative analysis methods to provide insights into human attention and behavior in AR applications.

Together these works demonstrate a well studied area of visualization and human behavioral analysis. However, one area that these visualization techniques rarely accommodate is in human behavioral analysis with physiological measures. This is especially sparse in augmented reality tools where physiological measures are often paired with AR sensor suits to monitor an individual's activity on real world tasks.

2.3 Insights into human performance with fNIRS

Functional Near-Infrared Spectroscopy (fNIRS) provides physiological measurements through non-invasive tracking of brain activity by monitoring oxygenated and deoxygenated hemoglobin levels [17]. fNIRS are often used as a brain-computer interface (BCI) when movement and portability are paramount to the task being measured [29]. Notably, this is often the case in virtual reality, augmented reality, and real-world tasks. Human behavior understanding can be amplified through these concentration measurements by inferring cognitive workload states in conjunction with synchronous multimodal measurements of an individual's actions and tasks.

A key application of fNIRS is assessing cognitive workload, namely employing behavioral models to infer workload capacity from structured tasks. These models facilitate understanding across both laboratory and real-world settings, predicting cognitive states from hemoglobin concentration data [3, 4]. Research, including works by McKendrick et al., validates the cross-person and cross-task applicability of these models, demonstrating their significance in translating lab findings to practical environments [23, 26].

2.4 Human Behavior based on fNIRS

Multimodal data, synchronously collected with fNIRS-based cognitive workload, enriches the analysis of human behavior, guiding the design of more responsive and adaptive real-world systems. Mark et al. provides a comparison study which incorporates various brain-body measures to offer insights into cognitive processes over time [24]. Similarly, Yuksel et al. demonstrate how adjusting task difficulty based on cognitive load readings and behavioral measurements can significantly improve learning efficiency, as seen in their adaptive piano training program [43].

In high-stakes environments like aviation, fNIRS is often used for monitoring cognitive workload and fatigue, offering insights into pilot engagement and decision-making processes [3, 4]. Various studies [13, 25, 36, 41] highlight the role of fNIRS in evaluating pilot performance in varied scenarios, including real and simulated flights, thereby showcasing the modality's adaptability and effectiveness in critical applications. On top of this, many of these studies rely upon the need for multiple modes of synchronous sensor data in addition to fNIRS physiological measurements to understand human behavior.

Integrating fNIRS with other synchronous time series modalities enriches behavioral analysis, allowing for a nuanced understanding of human cognition. This prompts the need for visualization tools to assist behavioral specialists in interpreting complex interconnected time series

datasets, specifically tools which link physiological measurements to specific behaviors and decisions.

2.5 Measuring workload effects in AR environments

The vast majority of previous studies about cognitive workload effects in AR environments focus on measuring the impact of using an AR headset on the wearer's mental workload during a task [2, 22, 31], rather than measuring the cognitive workload of a person who just so happens to be performing the task in an AR environment. Caarvida [1] and AutoVis [18], for instance, provide tools to explore automotive test data, but do not draw correlations between cognitive load and performer actions and errors.

More recent AR studies involve interfaces that evolve as a function of the individual's workload state, requiring real-time and multimodal behavioral analysis [15, 23]. This brings a new set of challenges to visualization tools. We need tools that can generalize to many environments, run in real time, and visualize many synchronous streams of data. Galati, Schoppa, and Lu implement a visualization tool in their AR pipeline [15]. This tool features an interactive exploration of user movements with respect to raw fNIRS signals, allowing experts to compare and identify areas of cognitive activity in the raw signal. However, this tool is tailored to handle data from this specific study. Furthermore, it is visualizing raw fNIRS data instead of classified workload states. While this is useful for neuroscientists who want a better understanding of spatial brain data, it is not as helpful for the broader study of human behavior, particularly by AR system developers who may not have a neuroscience background.

To the best of our knowledge, there are no such generalized tools that aide human behavior specialists in analyzing these many synchronous streams of data for augmented reality systems. Such a tool will improve both the efficiency of analysis as well as the conclusions that can be drawn from human behavior data.

3 OCARINA DATASET

The Ocarina dataset, collected by NGC as part of the DARPA PTG program, consists of data from simulated UH-60V helicopter copilot sessions, totaling approximately 3TB. It encompasses data from 7 participants and 33 sessions. Each participant participated in one to eleven sessions. Every session corresponded to a specific task scenario, with each task comprising multiple non-sequential procedures that could be completed in different orders. After recording, these procedures were extracted from the mission logs and designated alphabetical names from "a" to "f". This subset of six procedures was chosen from the nine possible ones as they accounted for 98.5% of all procedure occurrences. Each unique task scenario is identified by a distinct "trial ID" within the dataset, except for trials 2, 10, and 23, that represent the same task.

Participants. Data collection for the Ocarina dataset involved 7 participants. Three participants had previous piloting experience. No pilot had direct experience with the specific UH-60V cockpit. Three participants were engineers with experience developing the software for the UH-60V cockpit. Each had varying levels of experience with the system, but all were familiar with the cockpit. Two of the engineers were highly-versed in the logic of the system and had directly developed several of its capabilities. The seventh participant was a computer science professor at a large North American university. In total, participants completed 47 flights.

Data Collection Protocol. Participants were seated in front of a physical recreation of the UH-60V cockpit, with mission computers that replicate flight systems and simulate flight routes and in-flight events. They were outfitted with recording devices, including an fNIRS neuroimaging system. Additionally, a Microsoft HoloLens 2 was placed atop the fNIRS device. The HoloLens collected audio, video, IMU data containing accelerometer, gyroscope, and magnetometer readings, eye tracking data consisting of 3-dimensional vectors for gaze origin positions and directions, and hand tracking data consisting of 26 joint points for each hand. Throughout the data collection process, participants performed full flights, comprising pre-flight and flight phases. To advance to the flight phase, participants needed to complete nine procedures

during the pre-flight phase. The mission computer logs recorded all physical interactions the participants made with the simulator during each trial.

fNIRS Workload Classification. Predictions of the participants' cognitive workloads, spanning working memory, attention, and perception, are derived from the fNIRS time series measurements of hemoglobin concentration. These measurements, captured at a frequency of 10 Hz, include raw light intensities, HbO (oxyhemoglobin), and HbR (deoxy-hemoglobin) concentrations. These measurements serve as inputs for dedicated classifiers; at each time-step, three mental states are specified: perception, attention, and workload. Each mental state is classified as either "optimal", "overload", or "underload" with an associated classification confidence (we elaborate upon the interpretation of these classes in the following paragraph). Each workload category has its own classifier, which runs concurrently during recording sessions. The classifiers are generalized mixed effects models trained on data gathered from a previous test bed study that showed cross-task and cross-participant transfer [26]. The performer's classified mental state is first base-lined before recording and is not shown to the performer during collection.

Working memory capacity (which we hereafter refer to as *memory*) is an individual's ability to retain and manipulate information during task execution. *Attention* pertains to the individual's capacity to concentrate on specific tasks selectively. *Perception* is the individual's ability to interpret stimuli, both visual and auditory. The Ocarina dataset categorizes these cognitive facets into three states: optimal, overload, and underload. An optimal state denotes a balanced cognitive load conducive to task performance. An overloaded state suggests a cognitive burden exceeding an individual's capacity, potentially impairing the incorporation of new information [40]. Conversely, an underloaded state indicates a cognitive engagement below the individual's capacity, which may result in diminished focus [40]. It is critical to monitor multiple synchronous information streams alongside cognitive state data, as the presence of an overloaded or underloaded state does not invariably correlate with decreased task performance.

4 METHOD

4.1 Domain Requirements

The design requirements of *HuBar* were defined during a year-long collaboration with researchers, who are coauthors of this paper, actively developing an end-to-end task guidance system to support pre-flight procedures. In addition, we conducted multiple interviews with data scientists who actively work in fNIRS data analysis and data visualization to validate our design choices.

- [R1] **Performer behavior overview.** The experts stressed the importance of having the ability to visualize all performers' behavior in a single view. This helps trainers categorize performer expertise based on their behaviors across sessions. Trainers want to identify performers, for example, who may need additional training. Additionally, the trainers would like to know specific procedures where a certain performer excels or struggles. Finally, the trainers would like to detect and investigate clusters of similar sessions or performers. We propose a combination of the Scatter Plot and Summary Matrix Views to tackle this requirement.
- [R2] **Aligning and comparing multiple sessions.** Visualizing and comparing multiple sessions, each with multiple attributes based on time series data, can be challenging when there is no implicit sequentiality (as is the case with the Ocarina dataset). A time-aligned view of the procedures, errors, workload information, and task phase information would help trainers discern important landmarks within a particular session. Furthermore, this would enable trainers to compare landmarks and timestamps across multiple sessions and performers. To tackle this problem, we propose the Event Timeline View that combines the various streams of time series data along a common time axis, enabling seamless comparison across sessions.
- [R3] **Compare fNIRS data across sessions and visualize correlations between fNIRS data and performer behavior.** The experts

stated they were interested in visualizing and comparing fNIRS data across different performers at different levels of granularity. They would like to investigate fNIRS summaries for subjects across all sessions, while also being able to drill down into a single session and make comparisons between sessions. Furthermore, experts were particularly interested in understanding correlations between the mental states (overload, underload, and optimal) of performers for each workload category (attention, perception, memory) and errors made during sessions. Finally, experts would like to know the correlations between mental states and specific procedures. To this end, we propose the Workload Aggregations in the Overview, along with detailed fNIRS data for individual sessions in the Event Timeline and Summary Matrix Views. Furthermore, we display the correlations between errors and the mental states for both sessions and performers.

- [R4] **Detailed visualization of performer behavior.** To uncover associations between performer behavior, fNIRS predictions, actions, and errors, trainers need to explore individual sessions in great detail. Trainers would greatly benefit from being able to analyze data from the IMU and gaze sensors in conjunction with the ego-centric video captured by the AR headset, as it would allow them to detect patterns and establish connections between the various data streams. To meet this requirement, we propose the Detail View, which includes interactive visualizations for IMU and gaze data linked to the session video.

4.2 Visualization Design

We employed the "overview-first, zoom and filter, details-on-demand" strategy [33] as our guiding principle while formulating the visual design, with the goal of ensuring a user-centric approach that facilitates efficient exploration and comprehension of the multimodal data associated with performer sessions. The resulting tool, *HuBar*, is composed of four linked interactive views: the Overview (Figure 1(A) and 1(B)), the Event Timeline View, the Summary Matrix View, and the Detail View (Figures 1(C), 1(D), and 1(E), respectively).

Overview

The Overview consists of two sub-views: the Scatter Plot View shown in Figure 1(A), which allows users to select sessions based on various features, and the Workload Aggregation View shown in Figure 1(B), which displays cognitive workload and session duration information based on user selection.

Scatter Plot. The Scatter Plot View shown in Figure 1(A) serves as the starting point for exploration in *HuBar*. The Scatter Plot View categorizes sessions to facilitate the identification and comparison of similar sessions or outliers [R1]. The user can adjust the scatter plot to represent only the performers' physical activity (IMU, gaze) or brain activity (fNIRS) by selecting the desired data stream. Below, we detail the process of transforming time series data into 2D scatter plot points. Each point in the plot represents a session, and different symbols represent either trials or subjects, depending upon user selection. The user can lasso-select these symbols, and remaining views will update accordingly. Users can also opt to display only the types of tasks which appear most frequently in the dataset (e.g. "top 10").

Brain and Physical Activity. The user can toggle whether the points in the Scatter Plot View represent IMU, gaze, or fNIRS data. Toggling to IMU or gaze enables the user to select sessions based on the performer's physical activity throughout the session; IMU data represents the performer's body movement, whereas gaze data represents the displacement of their visual attention. To compare these time series, we transform them into 2D vector representations. We chose to do this using a shapelet-based [39] technique due to its ease of use and robust implementation through the `tslearn` library [35]. Although this algorithm requires some preprocessing of the data, such as normalizing time series to the same length, our system is agnostic of the technique. Other approaches (e.g. TS2Vec [42]) could be substituted in cases where the normalization of the time series could hide important information about the sessions.

In contrast, the user may toggle the Scatter Plot View to show fNIRS time-series data, shifting the focus of the exploration to brain activity throughout the sessions. A similar process could be applied to generate the projection of points based on fNIRS data as was used for the IMU and gaze data. However, in the current implementation of *HuBar*, we transform the raw fNIRS signal using the workload classification models described in Section 3 before generating the 2D vectors that are ultimately rendered in the Scatter Plot View.

Workload Aggregation View. We showcase the proportion of time spent in each mental state (overload, optimal, and underload) across the three workload categories (memory, attention, and perception) aggregated by the selection made in the scatter plot view (Figure 1(B)). To convey the ordinality of these mental states, we employ a sequential red color scale where light red represents underload, a medium shade indicates optimal conditions, and dark red signifies overload ■■■.

Furthermore, we present the error contribution linked to each mental state across all categories for each group. This metric is crucial, as it presents the correlation between performers' errors and their respective mental states for the three categories. To display this information effectively, we opted for an aligned position against a common scale plot. This choice facilitates easy comparison and identification of data of interest while saving vertical space and reducing visual clutter in the overall system. We intentionally avoided using a bar chart-based plot to minimize potential confusion with the workload bar chart. To avoid overlapping from identical correlation scores, we adapt the scale accordingly. Given a selected group of sessions g_i , we estimate the error correlation using Pearson Correlation (PC) between e , the error duration, and s , the state duration of each workload for the three categories (optimal, overload and underload). The sample points for these variables were collected for each procedure measured in seconds. Last, we highlight the average session duration for the selected groups [R3].

Event Timeline View

In the Event Timeline View shown in Figure 1(C), we coalesce data from four different data streams recorded during performer sessions into a unified, time-aligned visualization for each selected session. These sessions are organized by trial ID or subject ID, as chosen in the Scatter Plot View. Duration is represented along the x-axis, beginning at zero for each session.

Task steps or procedures are visualized using horizontal bars that extend for the duration of each session. Segments within these bars are color-coded according to the ongoing procedure at the corresponding timestamp. We excluded shades of red from this color scale to prevent any conflict with the scale used for the workload variable. Furthermore, we have an error bar that employs black segments to indicate errors at their corresponding timestamps. Next, we have the workload bar, where segments illustrate the performer's mental states (underload, optimal, overload) for the chosen workload category. Furthermore, the model confidence score for its predicted mental state is depicted using a line within the bar graph. Finally, we have the task phase indicator, which may be used to group task steps or procedures (e.g. in the case of the Ocarina dataset, this is where we use "PF" and "FL" to denote the pre-flight and flight stages, respectively).

The rationale for aligning the various data streams along the time axis is multi-faceted. First, employing a unified time scale across all selected sessions facilitates convenient evaluation of their respective durations. Moreover, it allows *HuBar* users to compare the performers' mental states and the errors committed across different sessions. In addition to inter-session evaluation, the design facilitates intra-session evaluation by enabling *HuBar* users to promptly identify error occurrences and establish potential correlations between errors and the corresponding procedures, mental states, and flight phase [R2] [R3].

Consider the scenario where the *HuBar* user wants to investigate a particular session. To do this, they simply brush the Event Timeline View along the time axis. This updates the Summary Matrix View, which employs transparency and opacity to highlight the procedures involved in the brushed section. This also updates the Detail View to display egocentric video and sensor data corresponding to the brushed

timestamps, enabling *HuBar* users to see the pilot's perspective and sensor readings for the selected period.

Summary Matrix View

The interviewed experts showed great interest in comparing errors, mental states, and prevalence of procedures within a session as well as across sessions [R1]. However, due to the non-linear nature of the procedures performed in many tasks, it can be challenging to discern these nuances when the data is visualized sequentially. To address these challenges, we propose the Summary Matrix View (Figure 1(D)), which complements the Event Timeline View to give a more nuanced picture of performer data. It includes pie charts for every procedure, where chart radius corresponds to procedure prevalence. The pie charts are shaded black and gray to based on the proportion of errors (represented by the black slice) within the corresponding procedure. Pie charts are employed here specifically to communicate two proportions simultaneously: (1) the proportion of a particular procedure in the duration of a session and (2) the proportion of error within each procedure for the session. This allows the *HuBar* user to compare procedures and associated errors with different procedures for the same session (horizontally), as well as with the same procedure for different sessions (vertically) [R1].

In addition to the pie charts, we show the proportion of errors and the distribution of mental states for the chosen workload category for each session. The provided checkbox can be used to show or hide the error contribution for mental states within the selected workload category [R3]. Since this error correlation corresponds to the individual session, we used the regular PC to calculate it (similar to the Workload Aggregation View). *HuBar* users can select the desired category either through the dropdown in this view, or by clicking the corresponding category label in the Workload Aggregation View. Additionally, transparency is used to fade out the non-selected categories in the workload aggregations view. This design decision is intended to help *HuBar* users retain focus by reducing visual clutter. Furthermore, the pie charts provide an on-hover tooltip which displays the correlation between errors e and the mental states s within the corresponding procedure p (p is a vector where $p_i = 1$ if the procedure i is the procedure in question and $p_i = 0$ otherwise). We calculate these values using Partial Correlation [5]. Let r_{se} be the correlation between s and e ; r_{sp} , the correlation between s and p ; and r_{ep} , the correlation between e and p . The Partial Correlation is computed as:
$$r_{se,p} = \frac{r_{se} - r_{sp}r_{ep}}{\sqrt{(1-r_{sp}^2)(1-r_{ep}^2)}}.$$

Detail View

One of the major requirements expressed by the interviewed experts was the ability to investigate individual sessions and observe performer (e.g. pilot) actions in detail [R4]. The Detail View was designed to meet this requirement (Figure 1(E)). The video view plays the egocentric video from the performer's perspective corresponding to the brushed timestamps. Below this, we use line plots to visualize data from the IMU and gaze sensors, capturing the performer's body and eye motion over time. The *HuBar* user can switch between the variables corresponding to the IMU and gaze sensors using their respective dropdowns. Finally, the segmented bar graph depicts the mental states for the chosen workload category throughout the session. The time window brushed in the Event Timeline View is highlighted in all three visualizations within this Detail View. All three visualizations can be brushed, similar to the Event Timeline View. Moreover, the brushes are all synchronized with each other and with the video player, facilitating seamless navigation and exploration of sessions. Aligning the IMU and gaze data with the video, workload information, and procedures enables the *HuBar* user to identify procedures with high levels of human motion, establish associations between motion levels and mental workload as well as errors, and navigate to these regions of interest in the video by simple brushing [R4].

5 EVALUATION

In the aviation industry, pilots often experience mental states of overload or underload which can have immediate consequences such as

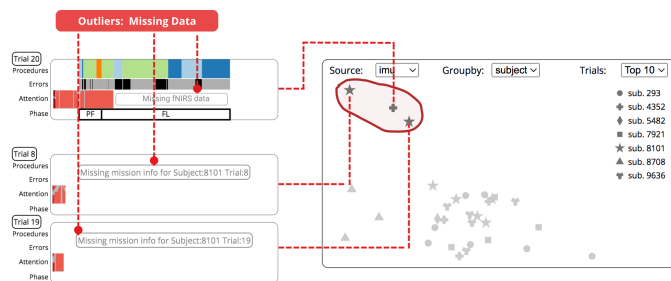


Fig. 2: Uncovering Data Quality Issues. On the right side, the scatterplot showcases sessions clustered by their IMU data, with glyphs encoding the subject ID. Variations in session counts per subject are evident, with some outliers identified in the upper left corner and highlighted through lasso selection. On the left side, the event timeline view reveals missing data points in trials 8, 19, and 20, likely attributed to mission log failures. Despite an initially comprehensive appearance, Trial 20 exhibited notable gaps in fNIRS data.

heightened stress, monotony, mental exhaustion, or fatigue. In addition to posing significant risks to flight safety, these short-term effects, if not addressed, can escalate into long-term issues such as psychosomatic or mental health disorders. In the following case studies, we describe how a pilot trainer and an AR guidance system developer can use *HuBar* to recognize and evaluate the factors contributing to overload or underload mental states in copilots.

5.1 Case Study 1: Unraveling the Triggers of Mental Underload in Copilots

To showcase how the *HuBar* system supports effective exploration of task sessions within real-world contexts, we present a case where a pilot trainer utilized the system. The trainer aimed to identify instances during flight procedures where a copilot might experience an underloaded mental state, discern potential causes behind such occurrences, and extract valuable insights from the data. The underload mental state is particularly concerning during a flight as it may indicate that the copilot is overly relaxed or not sufficiently focused.

Uncovering Data Quality in Flight Sessions. In any study focused on unraveling cognitive processes, data quality plays a critical role. Acknowledging this, the trainer began the analysis by utilizing the Scatterplot View to visualize the collected data across multiple sessions. For this particular task, she organized the data by trial, seeking to identify sessions where the underloaded mental state predominated. Analysis of the scatterplot first enabled her to identify outliers and anomalies that could indicate data quality issues, such as sensor failures or inaccuracies in data collection (see the right side of Figure 2). The trainer investigated these outliers using the Event Timeline View, which provided a detailed breakdown of data acquired throughout the sessions. As shown on the left side of Figure 2, this examination revealed missing data points in Trials 8, 19, and 20: Trials 8 and 19 only contained fNIRS data, and lacked crucial information like procedures and errors, likely due to mission log failures. Meanwhile, though Trial 20 appeared comprehensive at first glance, it exhibited notable gaps in fNIRS data, implying potential technical glitches or inconsistencies in recording procedures. After identifying the sessions with potential issues, the trainer opted to analyze a different cluster of sessions for further examination.

Understanding the Link Between Errors and Underloaded Mental States. Acknowledging that errors often signify underlying issues, the trainer scrutinized the Workload Aggregation View, concentrating on sessions displaying a notable correlation between errors and underloaded mental states using the error contribution plot. As shown in Figure 3, only one session (Trial 13) out of 10 sessions exhibited significant correlations. This implies that sessions falling under this trial demonstrate a strong association between the underloaded mental state and errors. Based on these findings, the trainer selected Trial 13 for deeper analysis.



Fig. 3: Workload Aggregation View. On the left side, a selection of sessions is made within the scatterplot to identify instances where the underloaded mental state prevails. On the right side, sessions are organized by trials, depicting workload and error contribution associated with each mental state across all categories for every trial group. Notably, Trial 13 reveals a substantial correlation between errors and underload state, as indicated by the prominent pink marker near the value of 1.

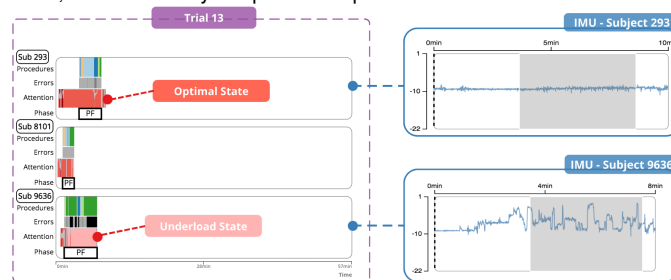


Fig. 4: Event Timeline and Detail Views. On the left side, the Event Timeline View presents sessions belonging to Trial 13, conducted by different subjects. Subject 293 demonstrates sustained optimal attention and minimal errors, while Subject 9636 encounters numerous errors under an underload attention state. On the right side, the Detail View displays IMU data for Subjects 293 and 9636, revealing distinct patterns in linear acceleration. Subject 293 exhibits consistent, controlled motion, while Subject 9636 shows considerable variation, indicative of frequent stops and starts.

Understanding Copilot Expertise Disparities through Motion Analysis and Error Correlation. In Trial 13, the trainer notes a correlation between performer errors and the underloaded mental state, with a correlation coefficient very close to 1 (see Figure 3). Upon transitioning to the Event Timeline View to analyze the sessions, the trainer quickly discovers that Trial 13 comprises very short sessions, specifically a task duration of under 10 minutes per subject (as shown in Figure 4). Examination of the phase feature in the Event Timeline View reveals that all sessions exclusively included the preflight phase (PF), explaining their brevity. Further scrutiny reveals that the tasks in the sessions performed by Subject 293 and Subject 9636 were completed in approximately 9 minutes. However, Subject 293 predominantly maintained an optimal attention workload state and exhibited relatively few errors, while Subject 9636 encountered numerous errors and primarily operated under an underload attention state. To delve deeper into this discrepancy, the pilot trainer navigates the Event Timeline View, brushing over the entire session for Subject 293 and subsequently moves to the Detail View to assess human motion using IMU data (see the right side of Figure 4). Notably, Subject 293's linear acceleration plots demonstrate consistent, controlled motion, contrasting with Subject 9636's plots, which exhibit considerable variation, suggesting frequent stops and starts. This disparity leads to the hypothesis that human motion correlates with the copilot's expertise level. To validate this conjecture, the pilot trainer reviews videos for each session, confirming her hypothesis. In the videos, it becomes evident that both subjects have a manual in front of them, but Subject 293 appears less reliant on it, whereas Subject 9636 frequently pauses to flip through the manual. This observation aligns with the notion that individuals less familiar with the task are prone to more errors and increased reliance on reference materials.

5.2 Case Study 2: Enhancing AR Guidance Systems through User Analysis

To showcase how *HuBar* facilitates the advancement of AR guidance system development, we present two scenarios wherein an AR guidance system developer uses the platform.

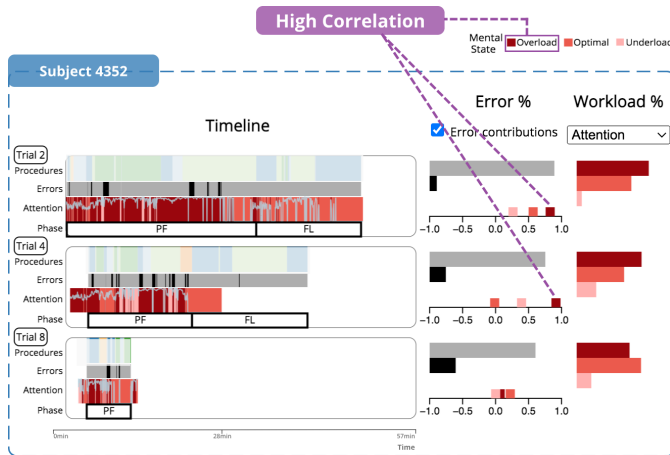


Fig. 5: The Event Timeline View displays sessions conducted by Subject 4352 alongside error and workload summaries. The workload summary reveals a consistent overload mental state across all sessions, notably correlating with errors, particularly in Trials 2 and 4, regardless of task variations.

Leveraging User Profiles to Optimizing AR Flight Guidance. Understanding end-users' characteristics is paramount for effective guidance system development. This example delves into the correlation between mental states and user characteristic profiles, emphasizing the importance of tailoring guidance measures to assist specific user groups. To achieve this, the AR guidance system developer aims to identify emerging patterns based on pilots' performance across various tasks. Unlike the previous case study, this one focuses on the overloaded mental state, rather than the underloaded one. The developer first groups the data by subject, assuming the issue stems from user profiles rather than the tasks themselves triggering mental states. The developer identifies Subjects 4352 and 293 as having significant error contributions to the overloaded mental state, despite both having previous piloting experience. Examining the Event Timeline View, the developer notes that Subject 293 and Subject 4352 completed five and three flights, respectively. Further investigation reveals that while Subject 293 displays higher overall percentages of the overload mental state during task performance, this condition is predominant in only one out of their five sessions, indicating variability in performance. Conversely, Subject 4352 consistently experiences overload across all sessions, despite task variations (see Figure 5). Furthermore, upon examining the correlations between errors and mental states in each session conducted by Subject 4352, it becomes evident that the overload mental state exhibits a strong correlation with errors, as shown in Figure 5. Examining their profiles further, Subject 293 emerges as a pilot with recent flight experience, having flown the most flights among their cohort, while Subject 4352 has been inactive in flying for 20 years. This underscores the need to consider user profiles in designing AR flight guidance systems, specifically different system versions or adaptive features tailored to individual user profiles.

Improving Performance and Mental State in AR Flight Guidance Systems. Consider an AR guidance system developer who sets out to evaluate the progression of novice engineers over multiple flight tasks, aiming to discern the factors underlying improvement and refine guidance mechanisms to minimize errors. The developer focuses on Subject 9636, a novice engineer, who performed the same flight task under normal conditions three times: Trial 2, Trial 10, and Trial 23, in sequential order. The Event Timeline View shows Subject 9636 consistently encountered challenges during the preflight phase across all trials (as shown in Figure 6). However, due to the sporadic nature of errors, pinpointing the specific procedures where the copilot struggled the most proved to be challenging. Further analysis through the Summary Matrix View revealed a consistent execution of tasks by Subject 9636 across sessions, with the most time spent on Procedure C during each session. Notably, significant errors were observed in Procedures A, D, and E during the first attempt (Trial 2). For Procedure E, the tooltip

visualization reveals a significant correlation (0.97) between errors and the overload mental state. Subsequent trials displayed improvement, particularly in Procedures A and E during the second attempt (Trial 10), where errors notably diminished, especially in Procedure E, dropping from over 70% to zero. However, errors emerged in Procedure F during this trial. This trend persisted in the third attempt (Trial 23), with a decline in performance in Procedure F but improvements in other procedures. Examining the Event Timeline View provided insights into the correlation between errors in Procedure F and the transition from pre-flight to flight phase, suggesting the necessity for additional guidance during this phase. Furthermore, analyzing the copilot's mental state through workload summaries revealed positive impacts with improved performance. Despite high levels of underload mental state during the first attempt (Trial 2), subsequent trials witnessed a decrease in underload mental state, albeit accompanied by an increase in overload mental state during the second attempt (Trial 10). By the third attempt (Trial 23), the copilot achieved minimal deviations from the optimal mental state. These findings emphasize the interplay between overcoming flight errors and improved copilot mental state. The developer acknowledges the imperative to focus efforts on enhancing guidance during the transition from preflight to flight to not only mitigate errors but also optimize the copilot's mental state. This case study underscores the iterative nature of analysis and adaptation essential in optimizing AR guidance systems for novice engineers' in-flight tasks.

5.3 Expert Interview

To validate our design decisions, we conducted a second round of interviews with five domain experts: three human factors and fNIRS experts (E1, E2, and E5), one biovisualization expert (E3), and one neuroinformatics algorithmic design expert (E4). All of them have experience with AR-enabled applications. Four of the experts had not previously seen the tool in action (E1, E2, E4, E5), whereas only one expert (E3) was part of the group that had previously assisted in identifying the system requirements (Section 4.1). In the experiment, experts were asked to explore a group of sessions of their choice according to their interests. The fNIRS experts were specifically asked to utilize the system to gain insights into the mental states of copilots by subject. Additionally, the fNIRS experts were queried on how this tool could be integrated into their workflows to enhance efficiency. Note that fNIRS experts needed to manually synchronize different data sources, such as video and workload, to analyze this data before they had *HuBar*. The design and visualization experts, on the other hand, were instructed to use the tool to explore the data with the goal of evaluating its usability.

Each interview took 50 minutes, and began with an overview of the project and gathering relevant background information from the participant (5 minutes). Second, we presented our system, including a demonstration, to the participant, addressing any questions or concerns (20 minutes). Third, participants were given the opportunity to select sessions of their preference from the Ocarina dataset (Section 3) to explore within the system (15 minutes). Finally, we engaged the participants in a discussion, asking questions about their initial impressions of the tool, its functionalities and features, and its potential application to their workflows (10 minutes).

The participants were given the freedom to use *HuBar* and explore the available sessions. However, they were also tasked with completing three specific assignments based on the selected sessions: 1) Identify sessions demonstrating a high correlation between the overload mental state and errors, 2) Identify the most prevalent procedures within and across sessions, and 3) Utilize *HuBar* to interpret the sessions. They were instructed to speak while using the system, following a "think aloud" protocol. While the participant performed the task, an investigator took notes related to the actions performed. After completion, the participants filled out a questionnaire to express their impressions on the usability of the system. In this section, we describe the insights gathered by the participants.

Expert Insights

Data Quality Assessment. Participants also use *HuBar* to assess the quality of the data. In particular, E2 was very interested in ensuring that

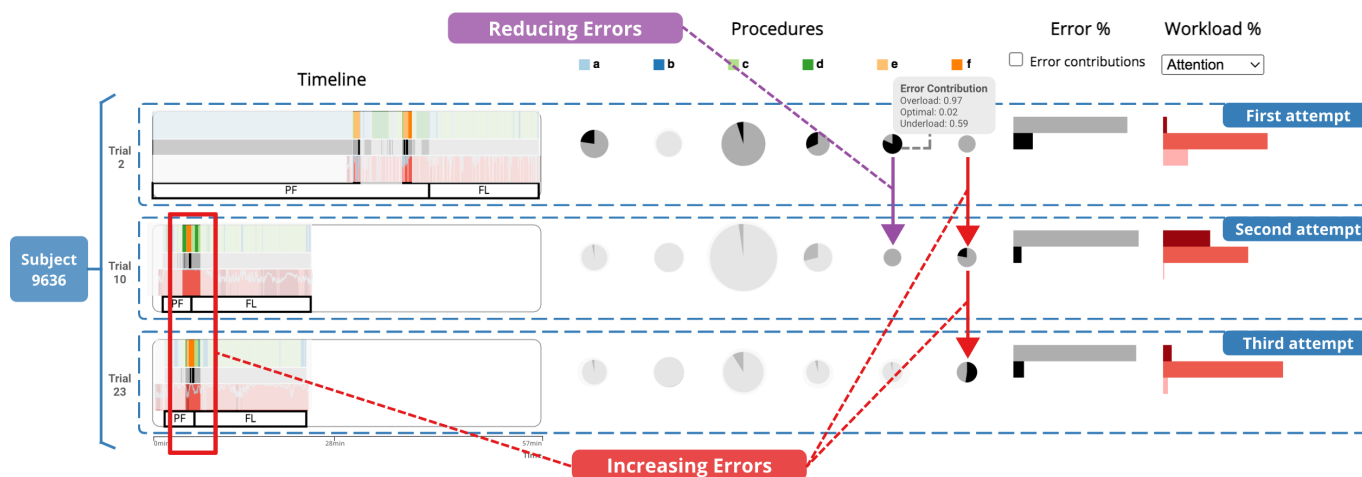


Fig. 6: Performance Overview for Subject 9636. The Timeline view, Matrix view, and Workload Summaries illustrate performance across three consecutive trials (same task conditions). Notable trends include consistent task execution, reduced errors (especially in Procedure E), increased errors in Procedure F that are correlated to the transition from the preflight (PF) to flight (FL) phase, and correlations between errors and mental states, particularly in Procedure E during the first trial where the overload mental state is correlated with errors, as shown in the tooltip. Workload summaries indicate mental state improvements, with the last trial showing predominantly optimal states.

all sensors data were included before drawing any conclusions about copilot behaviors. He utilized the Scatterplot View for this purpose. He identified outliers, hypothesizing that these sessions might have issues. Subsequently, he moved to the Event Timeline View to inspect the data confirming his hypothesis. Crucial information about procedures, errors, and flight phases was missing for the selected sessions, with only fNIRS data present. He noted that such occurrences are common due to sensor failures and expressed a desire to utilize the tool to identify such issues using the Scatterplot and Event Timeline Views. Afterward, he selected another group of sessions to continue his analysis. On the other hand, E1 was not particularly focused on identifying issues in the data. While analyzing the overload mental state of Subject 4352, she observed that in trial 20, the copilot remained mostly under the optimal mental state, unlike other trials where the overload mental state was predominant. However, upon referring to the Event Timeline View, she noticed that more than half of the fNIRS data was missing for this trial. Consequently, she determined that this trial should not be included in the analysis. E3 followed a similar approach to E1.

Procedures Analysis. E1, E2, and E4 extensively used the Summary Matrix View to identify the predominant procedures within and across sessions for each subject. Their approach was straightforward and effective. In contrast, E3 attempted to extract this information using the Event Timeline View by comparing the duration of each procedure across the session. However, after a brief attempt, they switched to the Summary Matrix View and quickly identified the predominant procedures. E3 highlighted that the Summary Matrix View, with its normalized data across sessions, provided a clearer focus on procedures compared to the Event Timeline View. E5 took a different approach, primarily using the Summary Matrix View to identify key procedures but also extensively employing the Event Timeline View to observe the frequency of these procedures throughout the session, focusing on copilot performance during each procedure.

Understanding Human Behavior through Error Analysis. The majority of participants began their analysis by examining the error contribution plot located in the Workload Aggregation View, organizing the data by subjects rather than trials. For instance, E1 used this view to identify subjects exhibiting a predominant overload mental state across various trials. To interpret the subjects' mental states during these trials, she navigated through the Event Timeline View and delved into the Detail View. By using the Detail View, specifically the IMU signals, she observed a significant amount of human motion at the beginning of the session, transitioning to a phase characterized by consistent and controlled motion. Upon revisiting the Event Timeline View, she noted

that this pattern correlated with the occurrence of errors and the flight phase, leading her to hypothesize that the pre-flight phase might be a contributing factor to errors and subsequent overload mental states due to heightened stress levels.

Managing Multimodal Data. Most participants extensively explored the various modalities available in the *HuBar* tool. Notably, E3 and E4 emphasized the tool's capability to visualize different data sources, including events (such as procedures and errors), fNIRS, IMU, gaze, and video, all of which were seamlessly integrated and synchronized. Among these modalities, video emerged as the preferred choice for all experts, serving as a vital resource for session analysis. E5, in particular, heavily relied on video analysis. With a clear understanding of the conditions for each trial, E5 was keen on inspecting segments within the session where abnormal events, such as weather disturbances, occurred, to evaluate the copilot's performance. To identify these events, E5 also used the Event Timeline View to locate procedures throughout the session. Additionally, E5 heavily relied on the IMU signal, particularly the accelerometer signal, to pinpoint segments in the video characterized by significant variance. These variations, evident through peaks and valleys in the accelerometer signal, helped identify critical moments for closer examination.

Expert Feedback

The participants provided highly positive feedback, demonstrating interest in utilizing *HuBar* for their tasks and providing suggestions for enhancing the system. Following the think-aloud experiment, they were asked for any additional comments or suggestions. Here are some of their responses:

- E1 liked the design and interactive part of the tool. She highlighted the selection of colors to encode mental states: "I really liked that progression (colors), a lot of people use like green, red and yellow to represent those states. And I really prefer what you guys have done, which is like the light to the dark red. I think that makes way more sense.". She also liked the usage of pie charts: "I really liked the use of pie charts here. I am not usually a big fan of them, but I think that that's an appropriate place for them. So I was happy to see a good pie chart.". Regarding interactivity, she appreciated the synchronized behavior exhibited by all components: "I think it was both intuitive and user friendly. Being able to lasso on the scatter plots makes things really, really, really easy to capture like little clusters that you're more interested in. I liked the brushing. It was responsive on both sides of the screen (components), so I don't have to go back and forth between different sections (components) in order to look at

something else, or just to switch things around.”

- E1 also found the Event Timeline View and Detail View useful to compare and validate hypotheses: “The bottom section, where I could see everything in comparison, side by side, the IMU data with the overload/underload state and having the video there so that you’re able to validate what it is you’re seeing and why, you’re seeing it. I thought that was very useful.”
- E2 liked the usage of scatterplot to detect outliers: “The outlier detection, or the outlier capability in the upper left was kind of really powerful. I would maybe like to see that expanded from just IMU (gaze or fNIRS) data, and maybe look at other kind of outliers, or be able to group by other kinds of data up there. So that was really useful.” He also found the Event Timeline View very powerful: “being able to see the procedure with the error and the workload state on the timeline view in the lower left. That was, also, I think, very helpful, really powerful, to be able to see those 3 things stacked up against each other.”
- E3 found the system’s capability to identify correlations to be effective and useful: “I found myself working a lot with the time (timeline view), with the event sequence. Even though I know that you cannot directly compare. You know the procedures with each other because they have multiple options to do these things. It still showed me like, very well what’s correlated? In which procedure didn’t the error occur? And then, how was that correlated to the mental state. I think, the timeline helped a lot.” E3 also liked the usage of different modalities to interpret the data: “It was definitely cool to look into the video because you kind of wanna know what’s going on. The other things are kinda abstract, and that just helps to relate a little bit to the situation. It is good to really connect what was exactly happening.”
- E4 emphasized the Event Timeline View’s capability to enable detailed segment inspection through brushing, facilitating deeper analysis: “I like the most, was the ability to take like a section of a trial, and then like overlay that with the raw measurements of behavior such as the IMU and other markers, and the videos really nice to also see, like a raw behavior there a little bit, I really like that.” He also appreciated the system’s full interactivity: “. . . each panel seems to complement each other, which is nice. I like that. I like that all the panels are tied to one another, so you can select trials in one, and then it shows it updates all the other panels and shows you nice statistics. It seems like well thought out and smooth interface.”
- E5 wants to integrate *HuBar* in his workflow: “part of what I do is to go through these kind of videos. Having more of that data (*HuBar*’s features), would kind of allow me to jump to things easier . . . As soon as you did that (brush segments of the timeline view and synchronize this with the video), I was like, that’s I wish I had that earlier.” He also found *HuBar* helpful to compare different sessions: “when you’re trying to make sense of the data in your analysis, you know what you might find. For example, you know there is something significantly different between two people or something. This tool would allow you to kind of quickly drill down into what’s actually going on. Either cognitively or behaviorally. So yeah, it’s helpful.”
- Suggestions: E1 and E5 suggested enriching the Summary Matrix View, for example, by including the proportion of mental states within the pie chart associated with errors. E2 suggested support for real-time monitoring. E3 suggested the use of pattern detection to presort the sessions. Finally, E4 suggested displaying raw fNIRS data, such as an activation map for brain signals, along with the locations of fNIRS sensors.

5.4 Usability

We assessed the usability of *HuBar* using the System Usability Score (SUS) [9], a robust tool widely recognized for evaluating system interfaces [6]. A mean SUS score above 80 is in the fourth quartile and is acceptable. To compute the SUS, we administered a survey at the

conclusion of the second interview, prompting participants to complete the standard SUS questionnaire, grading each of the 10 statements on a scale from 1 (strongly disagree) to 5 (strongly agree). The SUS grades systems on a scale between 1 and 100, and our system obtained an average score of 87 ± 9.58 .

6 DISCUSSION AND CONCLUSION

We presented *HuBar*, a novel visual analytics tool tailored for summarizing and comparing task performance sessions in Augmented Reality (AR). By integrating time series data from fNIRS measurements, gaze, and IMU data with session logs and videos, *HuBar* enables users to explore performer behavior and cognitive workload at various levels of granularity. Through interactive visualizations *HuBar* reveals patterns and anomalies in task performance, such as human errors and workload fluctuations, and their correlations with task phases. These insights support post-hoc analysis, aiding developers in refining task guidance strategies and enhancing AR-based training environments.

We believe *HuBar* integrates seamlessly into the ecosystem of AR-enabled task guidance development by enabling developers to assess the impact of different design decisions on performer cognitive workload. For example, specific 3D interfaces designed to guide users through tasks can trigger variation in performer cognitive load depending on design. ARTiST [37], for instance, leverages this by proposing a text simplification approach to reduce performer cognitive load. In turn, *HuBar* could facilitate more detailed exploration of the actual impact of such systems on performers. This capability could help developers create more adaptable task guidance systems that customize instructions to the performer’s mental state.

Limitations. While *HuBar* enables users to understand the frequency and magnitude of performer movement through IMU and gaze data, it does not include an explicit representation summarizing spatial relationships between the performer and their surrounding environment. In other words, *HuBar* makes it easy to tell when the performer moves, but their exact pose and, in turn, action may not always be clear. This limitation persists even when IMU and gaze time series are analyzed in conjunction with egocentric video, as many AR headsets have a limited field of view which could leave important hand movements and interactions with the environment off-camera. Furthermore, *HuBar* does not include a visual representation of the raw data output by fNIRS sensors, instead opting for aggregated workload classification labels at each time step. While this approach enhances data interpretability for a broader audience, it may occlude anomalies in sensor performance or details that could be of interest to a brain data expert. This also creates blind reliance on the workload classifiers, with limited ability to identify potential classification errors.

Future work. First, we plan to conduct a larger user study with participants from different backgrounds to understand how well our design can adapt to new users. To intervene promptly in response to emerging issues or fluctuations in cognitive workload, we plan to enable real-time monitoring of task performance sessions. To aid in better data quality assessment and model interpretation, we also plan to explore scalable visual metaphors for analyzing fNIRS raw time series data, which may be composed of up to several dozen streams. This raw data will enable users to understand how different brain parts respond to specific stimuli and note data quality issues. On the machine learning front, we would like to explore techniques to automate the detection of relevant patterns and anomalies within task performance data [20]. This may include developing algorithms to classify human errors and identify optimal task guidance strategies based on historical data. Finally, we have primarily explored the aviation domain in our use cases due to the availability of relevant data, but it is important to note that our tool is applicable across various domains, which we plan to explore in future work, as our methods work with any multimodal time series data.

ACKNOWLEDGMENTS

This work was supported by the DARPA PTG program. Any opinions, findings, and conclusions or recommendations expressed in this mate-

rial are those of the authors and do not necessarily reflect the views of DARPA. Erin McGowan is funded by an NYU Tandon Future Leader Fellowship.

REFERENCES

- [1] A. Achberger, R. Cutura, O. Türksoy, and M. Sedlmair. Caarvida: Visual Analytics for Test Drive Videos. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces, AVI '20*. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3399715.3399862 3
- [2] H. Atici-Ulusu, Y. D. İkiz, O. Taskapilioglu, and T. Gunduz. Effects of augmented reality glasses on the cognitive load of assembly operators in the automotive industry. *International Journal of Computer Integrated Manufacturing*, 34(5):487–499, May 2021. doi: 10.1080/0951192X.2021.1901314 3
- [3] H. Ayaz, B. Onaral, K. Izzetoglu, P. A. Shewokis, R. McKendrick, and R. Parasuraman. Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: empirical examples and a technological development. *Frontiers in Human Neuroscience*, 7, 2013. doi: 10.3389/fnhum.2013.00871 3
- [4] H. Ayaz, P. A. Shewokis, S. Bunce, K. Izzetoglu, B. Willems, and B. Onaral. Optical brain monitoring for operator training and mental workload assessment. *Neuroimage*, 59(1):36–47, 2012. Publisher: Elsevier. 2, 3
- [5] K. Baba, R. Shibata, and M. Sibuya. Partial Correlation and Conditional Correlations as Measures of Conditional Independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, Dec. 2004. doi: 10.1111/j.1467-842X.2004.00360.x 5
- [6] A. Bangor, P. T. Kortum, and J. T. Miller. An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6):574–594, July 2008. doi: 10.1080/10447310802205776 9
- [7] R. Beams, E. Brown, W.-C. Cheng, J. S. Joyner, A. S. Kim, K. Kontson, D. Amiras, T. Baeuerle, W. Greenleaf, R. J. Grossmann, A. Gupta, C. Hamilton, H. Hua, T. T. Huynh, C. Leuze, S. B. Murthi, J. Penczek, J. Silva, B. Spiegel, A. Varshney, and A. Badano. Evaluation Challenges for the Application of Extended Reality Devices in Medicine. *Journal of Digital Imaging*, 35(5):1409–1418, Oct. 2022. doi: 10.1007/s10278-022-00622-x 2
- [8] D. Bohus, S. Andrist, A. Feniello, N. Saw, M. Jalobeanu, P. Sweeney, A. L. Thompson, and E. Horvitz. Platform for Situated Intelligence, Mar. 2021. arXiv:2103.15975 [cs]. doi: 10.48550/arXiv.2103.15975 2
- [9] J. Brooke. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*. CRC Press, 1996. Num Pages: 6. 9
- [10] S. Castelo, J. Rulff, E. McGowan, B. Steers, G. Wu, S. Chen, I. Roman, R. Lopez, E. Brewer, C. Zhao, J. Qian, K. Cho, H. He, Q. Sun, H. Vo, J. Bello, M. Krone, and C. Silva. ARGUS: Visualization of AI-Assisted Task Guidance in AR. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1313–1323, Jan. 2024. doi: 10.1109/TVCG.2023.3327396 2
- [11] G. Y.-Y. Chan, L. G. Nonato, A. Chu, P. Raghavan, V. Aluru, and C. T. Silva. Motion Browser: Visualizing and Understanding Complex Upper Limb Movement Under Obstetrical Brachial Plexus Injuries. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):981–990, Jan. 2020. doi: 10.1109/TVCG.2019.2934280 3
- [12] DARPA. Perceptually-enabled Task Guidance. <https://www.darpa.mil/program/perceptually-enabled-task-guidance>. 2
- [13] F. Dehais, A. Dupres, G. Di Flumeri, K. Verdier, G. Borghini, F. Babiloni, and R. Roy. Monitoring Pilot's Cognitive Fatigue with Engagement Features in Simulated and Actual Flight Conditions Using an Hybrid fNIRS-EEG Passive BCI. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 544–549, Oct. 2018. ISSN: 2577-1655. doi: 10.1109/SMC.2018.00102 3
- [14] B. D. Fulcher. Feature-Based Time-Series Analysis. In *Feature Engineering for Machine Learning and Data Analytics*. CRC Press, 2018. Num Pages: 30. 3
- [15] A. Galati, R. Schoppa, and A. Lu. Exploring the SenseMaking Process through Interactions and fNIRS in Immersive Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2714–2724, May 2021. doi: 10.1109/TVCG.2021.3067693 3
- [16] D. Gong, G. Medioni, and X. Zhao. Structured Time Series Analysis for Human Action Segmentation and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1414–1427, July 2014. doi: 10.1109/TPAMI.2013.244 3
- [17] M. Izzetoglu, K. Izzetoglu, S. Bunce, H. Ayaz, A. Devaraj, B. Onaral, and K. Pourrezaei. Functional near-infrared neuroimaging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(2):153–159, June 2005. doi: 10.1109/TNSRE.2005.847377 3
- [18] P. Jansen, J. Britten, A. Häusele, T. Segschneider, M. Colley, and E. Rukzio. AutoVis: Enabling Mixed-Immersive Analysis of Automotive User Interface Interaction Studies. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3580760 3
- [19] T. Jiang, D. Yu, Y. Wang, T. Zan, S. Wang, and Q. Li. HoloLens-Based Vascular Localization System: Precision Evaluation Study With a Three-Dimensional Printed Model. *Journal of Medical Internet Research*, 22(4):e16852, Apr. 2020. doi: 10.2196/16852 2
- [20] M. Kloska, G. Grmanova, and V. Rozinajova. Expert enhanced dynamic time warping based anomaly detection. *Expert Systems with Applications*, 225:120030, Sept. 2023. doi: 10.1016/j.eswa.2023.120030 9
- [21] L. Liu, Y. Peng, M. Liu, and Z. Huang. Sensor-based human activity recognition system with a multilayered model using time series shapelets. *Knowledge-Based Systems*, 90:138–152, Dec. 2015. doi: 10.1016/j.knsys.2015.09.024 3
- [22] C. Maag, N. Schömig, F. Naujoks, I. Karl, A. Keinath, and A. Neukum. Measuring workload effects of augmented reality head-up displays using detection response task. *Transportation Research Part F: Traffic Psychology and Behaviour*, 92:201–219, Jan. 2023. doi: 10.1016/j.trf.2022.11.010 3
- [23] F. Maitz, L. Ribeiro Skreinig, D. Kalkofen, and S. C. Wriessnegger. Towards Neuroadaptive Augmented Reality Piano Tutorials. In *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE)*, pp. 450–455, Oct. 2023. doi: 10.1109/MetroXRINE58569.2023.10405829 3
- [24] J. A. Mark, A. Curtin, A. E. Kraft, M. D. Ziegler, and H. Ayaz. Mental workload assessment by monitoring brain, heart, and eye with six biomedical modalities during six cognitive tasks. *Frontiers in Neuroergonomics*, 5, Mar. 2024. Publisher: Frontiers. doi: 10.3389/fnrgo.2024.1345507 3
- [25] J. A. Mark, A. E. Kraft, M. D. Ziegler, and H. Ayaz. Neuroadaptive Training via fNIRS in Flight Simulators. *Frontiers in Neuroergonomics*, 3, Mar. 2022. Publisher: Frontiers. doi: 10.3389/fnrgo.2022.820523 3
- [26] R. McKendrick, B. Feest, A. Harwood, and B. Falcone. Theories and Methods for Labeling Cognitive Workload: Classification and Transfer Learning. *Frontiers in Human Neuroscience*, 13, Sept. 2019. Publisher: Frontiers. doi: 10.3389/fnhum.2019.00295 3, 4
- [27] V. Mollyn, R. Arakawa, M. Goel, C. Harrison, and K. Ahuja. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3581392 2
- [28] A. Nijholt. Towards Social Companions in Augmented Reality: Vision and Challenges. In N. A. Streitz and S. Konomi, eds., *Distributed, Ambient and Pervasive Interactions. Smart Living, Learning, Well-being and Health, Art and Creativity*, Lecture Notes in Computer Science, pp. 304–319. Springer International Publishing, 2022. doi: 10.1007/978-3-031-05431-0_21 2
- [29] P. Pinti, I. Tachtsidis, A. Hamilton, J. Hirsch, C. Aichelburg, S. Gilbert, and P. W. Burgess. The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1464(1):5–29, 2020. doi: 10.1111/nyas.13948 3
- [30] B. Puladi, M. Ooms, M. Bellgardt, M. Cesov, M. Lipprandt, S. Raith, F. Peters, S. C. Möhlhenrich, A. Prescher, F. Hölzle, T. W. Kuhlen, and A. Modabber. Augmented Reality-Based Surgery on the Human Cadaver Using a New Generation of Optical Head-Mounted Displays. *JMIR Serious Games*, 10(2):e34781, 2022. doi: 10.2196/34781 2
- [31] Y. Qin and T. Bulbul. An EEG-Based Mental Workload Evaluation for AR Head-Mounted Display Use in Construction Assembly Tasks. *Journal of Construction Engineering and Management*, 149(9):04023088, Sept. 2023. Publisher: American Society of Civil Engineers. doi: 10.1061/JCEMD4.COENG-13438 3
- [32] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.-K. R. Choo. Imaging and fusing time series for wearable sensor-based human activity recognition. *Information Fusion*, 53:80–87, Jan. 2020. doi: 10.1016/j.inffus.2019.06.014 3
- [33] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy

- for Information Visualizations. In B. B. Bederson and B. Shneiderman, eds., *The Craft of Information Visualization*, Interactive Technologies, pp. 364–371. Morgan Kaufmann, San Francisco, Jan. 2003. doi: 10.1016/B978-155860915-0/50046-9 4
- [34] E. T. Solovey, A. Girouard, K. Chauncey, L. M. Hirshfield, A. Sassaroli, F. Zheng, S. Fantini, and R. J. Jacob. Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, UIST '09, pp. 157–166. Association for Computing Machinery, New York, NY, USA, Oct. 2009. doi: 10.1145/1622176.1622207 2
- [35] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods. Tslern, A Machine Learning Toolkit for Time Series Data. *Journal of Machine Learning Research*, 21(118):1–6, 2020. 4
- [36] K. J. Verdière, R. N. Roy, and F. Dehais. Detecting Pilot's Engagement Using fNIRS Connectivity Features in an Automated vs. Manual Landing Scenario. *Frontiers in Human Neuroscience*, 12, Jan. 2018. Publisher: Frontiers. doi: 10.3389/fnhum.2018.00006 3
- [37] G. Wu, J. Qian, S. Castelo Quispe, S. Chen, J. Rulff, and C. Silva. ARTiST: Automated Text Simplification for Task Guidance in Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642772 9
- [38] T. L. Xu, K. de Barbaro, D. H. Abney, and R. F. A. Cox. Finding Structure in Time: Visualizing and Analyzing Behavioral Time Series. *Frontiers in Psychology*, 11, 2020. doi: 10.3389/fpsyg.2020.01457 3
- [39] L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pp. 947–956. Association for Computing Machinery, New York, NY, USA, 2009. doi: 10.1145/1557019.1557122 4
- [40] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock. State of science: mental workload in ergonomics. *Ergonomics*, 58(1):1–17, Jan. 2015. doi: 10.1080/00140139.2014.956151 4
- [41] J. Yuan, X. Ke, C. Zhang, Q. Zhang, C. Jiang, and W. Cao. Recognition of Different Turning Behaviors of Pilots Based on Flight Simulator and fNIRS Data. *IEEE Access*, 12:32881–32893, 2024. doi: 10.1109/ACCESS.2024.3367447 3
- [42] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu. TS2Vec: Towards Universal Representation of Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8980–8987, June 2022. Number: 8. doi: 10.1609/aaai.v36i8.20881 4
- [43] B. F. Yuksel, K. B. Oleson, L. Harrison, E. M. Peck, D. Afergan, R. Chang, and R. J. Jacob. Learn Piano with BACH: An Adaptive Learning Interface that Adjusts Task Difficulty Based on Brain State. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 5372–5384. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2858036.2858388 3
- [44] S. Öney, N. Pathmanathan, M. Becher, M. Sedlmair, D. Weiskopf, and K. Kurzhals. Visual Gaze Labeling for Augmented Reality Studies. *Computer Graphics Forum*, 42(3):373–384, 2023. doi: 10.1111/cgf.14837 3