Available online at www.sciencedirect.com
ScienceDirect
Interacting with Computers 19 (2007) 314–329
Interacting with Computers
www.elsevier.com/locate/intcom

TAPRAV: An interactive analysis tool for exploring workload aligned to models of task execution

Brian P. Bailey *, Chris W. Busbey, Shamsi T. Iqbal

Department of Computer Science, 201 N. Goodwin Avenue, University of Illinois, Urbana, IL 61801, USA

Received 11 April 2006; received in revised form 11 January 2007; accepted 18 January 2007
Available online 26 January 2007

Abstract

Pupillary response is a valid indicator of mental workload and is being increasingly leveraged to identify lower cost moments for interruption, evaluate complex interfaces, and develop further understanding of psychological processes. Existing tools are not sufficient for analyzing this type of data, as it typically needs to be analyzed in relation to the corresponding task's execution. To address this emerging need, we have developed a new interactive analysis tool, TAPRAV. The primary components of the tool include; (i) a visualization of pupillary response aligned to the corresponding model of task execution, useful for exploring relationships between these two data sources; (ii) an interactive overview + detail metaphor, enabling rapid inspection of details while maintaining global context; (iii) synchronized playback of the video of the user's screen interaction, providing awareness of the state of the task; and (iv) interaction supporting discovery driven analysis. Results from a user study showed that users are able to efficiently interact with the tool to analyze relationships between pupillary response and task execution. The primary contribution of our tool is that it demonstrates an effective visualization and interaction design for rapidly exploring pupillary response in relation to models of task execution, thereby reducing the analysis effort.

Keywords: Mental workload; Pupil size; Task models; Visualization

1. Introduction

Pupillary response is a reliable indicator of mental workload and is being increasingly leveraged in many areas of research, e.g., to identify lower cost moments for task interruption (Iqbal et al., 2005), evaluate complex interfaces (Marshall, 2003), and develop further understanding of psychological processes (Schluroff et al., 1986). It is well established that relative increases in a user's pupil size, or *pupillary response*, has a positive correlation with increases in their mental processing effort, or *workload* (Backs and Walrath, 1992; Beatty, 1982; Granholm et al., 1996; Just et al., 2003; Verney et al., 2001).

For example, in our prior work (Iqbal et al., 2005), we leveraged the use of pupillary response to better understand where periods of lower mental workload occur during execution of interactive tasks, as delivering information at these moments could reduce the cost of interruption (Miyata and Norman, 1986). For several tasks, we developed models of their execution by decomposing them into their component goals and operators (Card et al., 1983). In a lab setting, users performed the tasks while their pupil size was measured using an eye tracking system. By analyzing the cursor and gaze cues in the interaction videos, we were able to determine the start and end timestamps for each subtask in the model, allowing it to be precisely aligned to the pupillary response data. The task model and pupil data were then entered into a spreadsheet application for analysis.

However, prior to conducting formal statistical analysis, we first wanted to employ visualization techniques and retrieve descriptive statistics in order to make sense of and explore relationships between these data sources, e.g., to quickly determine a user's workload at specific

* Corresponding author. Tel.: +1 217 333 6106; fax: +1 217 244 6869.
E-mail address: bpbailey@uiuc.edu (B.P. Bailey).

0953-5438/$ - see front matter
doi:10.1016/j.intcom.2007.01.004

parts of the task, identify macro-level patterns in the workload data, and locate any unexpected areas of potential interest. This would allow us to informally generate, test, and refine hypotheses and gain further confidence in later analysis results.

Unfortunately, existing analysis tools such as GazeTracker, Data Viewer, and SigmaPlot are not sufficient for exploring this type of data. The foremost problem is that graphs generated with these tools do not allow interactive analysis of the pupillary response data in relation to the model of task execution. For example, descriptive statistics for the response data cannot be retrieved at different locations within the task model, which is one of the most common and important analysis needs. Due to high-resolution sampling of the pupil (e.g., up to 250 Hz) and use of a hierarchical model, the data also needs to be explored at various levels of detail. As a result, we had to generate many related, but independent graphs; repeatedly switch between the graphs and numeric spreadsheets; and invest a large effort into programming complex macros to sort, filter, and compare different parts of the data sets. These limitations severely inhibited our ability to explore and understand relationships between pupillary response and task execution.

Though this analysis scenario was grounded in our prior work, the need to analyze pupillary response in relation to models of task execution is in fact much broader. For example, pupillary response has been leveraged in numerous controlled experiments to study how task complexity relates to psychological complexity (Backs and Walrath, 1992; Beatty, 1982; Granholm et al., 1996; Hyona et al., 1995; Schluroff et al., 1986; Verney et al., 2001). Pupillary response has also been proposed as a new metric by which to evaluate complex user interfaces (Marshall, 2003). Anyone conducting these types of experiments or proposed evaluations would follow a similar methodology and would have analysis needs similar to those previously described.

In this work, we describe the design, use, and evaluation of a new interactive analysis tool, TAPRAV (Task Aligned Pupillary Response Analysis and Visualization). Our work drew upon several existing techniques for exploring other types of high-resolution temporal data, e.g., see (Casares et al., 2002; Mills et al., 1992; Stolte et al., 1999), and applied them to produce an effective tool that facilitates interactive analysis of pupillary response data aligned to a hierarchical model of task execution.

The main components of our tool include (i) a visualization of pupillary response aligned to the corresponding model of task execution, useful for making sense of relationships between these two data sources; (ii) an interactive overview + detail metaphor, enabling rapid inspection of specific parts of the aligned data while maintaining global context; (iii) synchronized playback of the video of the user's screen interaction, allowing awareness of the current state of the task; and (iv) interaction supporting discovery driven analysis; including interactive retrieval of descriptive statistics for pupillary response within any part of the task model, marking points of interest, creating multiple views on the data, recording analysis notes, and navigating data sets for multiple users. Results from a user study showed that users can efficiently interact with the tool to analyze relationships between pupillary response and task execution for ecological data sets.

1.1. Contributions

The principal contribution of TAPRAV is that it demonstrates an effective visualization and interaction design for rapidly analyzing pupillary response data in relation to hierarchical models of task execution, reducing the effort required to analyze this type of data. By lowering the analysis burden, which is currently very high, our tool can facilitate broader use of this type of physiological data in both research and practice. Also, our tool leverages the basic visualization metaphor of overview + detail, but extends this metaphor to support exploration of a temporally aligned data set – pupillary response aligned to a corresponding model of task execution.

2. Related work

We describe measures of mental workload and task models along with several projects linking them together and their common analysis needs. We discuss why existing tools are not sufficient for addressing these needs and review visualizations influencing those used in our tool.

2.1. Measures of mental workload

Mental workload is generally accepted to be the ratio of attentional resources allocated to a task versus the total resources available (Moray et al., 1979). There are three categories of mental workload measures; subjective (Hart and Staveland, 1988), performance-based (Wickens, 2002), and physiological (Kramer, 1991). The advantage of physiological measures is that they are continuous, enabling access to ephemeral changes in a user's mental processing effort (Kramer, 1991).

Physiological measures include event-related potential (Kok, 1997), electroencephalogram (Schacter, 1977), heart rate variance (Rowe et al., 1998), and pupil size (Beatty, 1982; Pomplun and Sunkara, 2003). Our work to date has focused on the use of pupillary response, as it measures workload holistically, is low latency, and offers an immediate measure, i.e., a few samples relative to a baseline value indicates workload (Kramer, 1991). Most newer eye tracking systems are able to measure pupil size to about a hundredth of a millimeter and at very high sampling rates, e.g., up to 250 Hz.

The average human pupil is about 5 mm and increases in pupil size correlate with increases in task demands (Beatty, 1982). Task-evoked increases in pupil size are usually less than 20% above the baseline (Beatty, 1982). Eye tracking

systems typically log the raw pupil and gaze data to a file. To extract workload from this data, one must first process the eye data (e.g., filter eye blinks, correct for saccades, interpolate missed values, etc.) and compute the relative increase in pupil size over a baseline, typically recorded at the beginning of the experiment.

Rising interest in the use of pupillary response as a research instrument can be attributed to better understanding of how to interpret the raw data, availability of less physically intrusive hardware at lower cost, and the enduring need to measure workload in many controlled experiments (Kramer, 1991). Our tool currently supports the use of pupillary response as the measure of mental workload, but the visualization and interaction techniques demonstrated could apply to the design of similar analysis tools for other measures of workload.

2.2. Models of task execution

A model of task execution, or *task model*, represents the hierarchical decomposition of the execution structure of a task. Models can be developed using any number of well-known modeling techniques such as Hierarchical Task Analysis (Kirwan and Ainsworth, 1992), Event Perception Theory (Zacks and Tversky, 2001), and GOMS (Card et al., 1983; John, 1995; John and Kieras, 1996). For example, when applying GOMS, the goal structure of a task is recursively decomposed into its elementary perceptual, cognitive, and motor operators, though any level of detail is possible. Our work assumes that an analyst will apply an existing modeling technique to define the hierarchical and sequential structure of the experimental tasks. The resulting models must accurately reflect users' specific execution sequences realized during the experiment and be at a level of detail consistent with the types of research questions being posed.

Once developed, the models of task execution, along with their temporal alignment to the pupillary response data, can be specified for our tool using a relatively simple markup language. The language allows the hierarchical and sequential execution structure of a task to be specified, which is sufficient for describing most of the tasks used in controlled experiments involving pupillary response, e.g., see tasks used in (Backs and Walrath, 1992; Iqbal et al., 2004, 2005; Schluroff et al., 1986; Verney et al., 2001, 2004). The execution sequences of these tasks must typically be tightly controlled in order for the corresponding pupillary response data to be properly compared across users. However, future work could explore extensions to the language and corresponding visualization that would support temporal overlap between elementary operators, as in CPM-GOMS (John et al., 2002), or that would support variable or more complex execution structures such as those supported by ConcurTaskTrees (Paternò et al., 1997).

The task models can be produced prior to users' execution of tasks in an experiment, e.g., to study how workload changes in relation to the structure of a task (Iqbal et al., 2005); or produced after an experiment by analyzing traces of users' execution of the tasks, e.g., to study how workload affects a user's selection of input modalities (Oviatt et al., 2004).

2.3. Workload relative to task execution

Many projects have measured pupillary response during execution of tasks and explored their relationship. For example, in our own prior work on interruption management, we wanted to understand where moments of lower workload occur during task execution (Iqbal et al., 2005), as interrupting at these moments could reduce the cost of interruption (Miyata and Norman, 1986). For example, one common, objective measure of the cost of interruption is the amount of time needed to resume an interrupted primary task (Altmann and Trafton, 2004).

We developed hierarchical models for several tasks and had users perform those tasks while their pupillary response was continuously measured using an eye tracking system. The models were then temporally aligned to the response data and statistically analyzed, but this was a significant struggle without access to interactive tools for first exploring and making sense of relationships between these two data sources.

Pupillary response has been proposed as a new metric by which to evaluate complex interfaces (Marshall, 2003). The basic vision is that users would perform a series of tasks with one or more interface designs while their pupillary response was measured. Designers would then align the response data to the models of task execution and either select the design that imposes the least workload overall or re-design areas within a specific design that have unacceptably high workload. Our work contributes to realizing this vision by providing a usable software tool that reduces the effort required for performing this type of analysis on pupillary response data.

In cognitive psychology, researchers have sought to understand the relationship between psychological processes and syntactic complexity of written language (Schluroff et al., 1986). To further explicate this nebulous relationship, users were asked to transform sentences with different levels of ambiguity while their pupil size was measured. The structure of the transformation process was then aligned to the pupillary response data and analyzed. Many similar experiments have been conducted to further understand the relationship between task complexity and psychological processes, e.g., see work in (Backs and Walrath, 1992; Granholm et al., 1996; Hyona et al., 1995; Takahashi et al., 2000; Verney et al., 2001).

Though not exhaustive, this sample of work offers strong evidence that pupillary response does in fact often need to be explored in relation to models of task execution. The benefit of using our tool is that it reduces the effort required to analyze relationships between these two data sources.

2.4. Tools for analyzing pupillary response and other behavioral data

We reviewed software analysis tools from two leading companies that sell eye tracking systems, GazeTracker from Applied Science Labs and Data Viewer from SR Research. Both tools offer support for analyzing pupil data. The tools can be used to generate graphs of pupillary response data over time, temporally zoom in/out of the data, and play videos of a user's screen interaction. However, the tools do not allow the response data to be *interactively* analyzed in relation to a task's execution, the zooming is disorienting as it does not maintain global context, and video playback is not synchronized to the other data sources. The lack of these features inhibits the ability to rapidly explore and understand relationships between the multiple data sources.

Our work seeks to develop a tool that overcomes these central limitations; allowing rapid, interactive analysis of pupillary response data in relation to models of task execution and allowing synchronized playback of the video of screen interaction. However, the particular visualization and interaction techniques used in our tool could be replicated within existing commercial tools.

In addition, numerous interactive visualization tools have been developed that allow behavioral data such as eye gaze, task completion time, and task execution traces to be explored in relation to prescribed models of task execution, e.g., see work by (Paganelli and Paternò, 2003) and the review in (Ivory and Hearst, 2001). Our work differs in that we are interested specifically in facilitating analysis of pupillary response data in relation to the hierarchical structure of a task's execution, which itself is a significant research problem, as discussed in the previous section.

2.5. Influential visualizations

One goal of our visualization was to show the pupillary response data visually aligned to the task model. The response data could be naturally graphed over time, but linking the task model was more challenging. We thus drew upon time-based visualizations resembling those used in typical scheduling charts as well as those used in CPM-GOMS (John et al., 2002). With this type of temporal visualization of the task model, we could visually align it to the pupillary response data.

Another important goal was to allow the data to be interactively explored at various levels of detail while maintaining global context. Studies show that maintaining context is an important factor when navigating large data sets (Baudisch et al., 2001; Hornbæk et al., 2002) and our early experiences have indicated that maintaining global context is also important when exploring details of high-resolution pupil data. For example, this would allow the analyst to inspect localized changes in workload as part of a surrounding trend in the data. One possible solution would be to use a distortion-based visualization such as a fisheye view (Sarkar and Brown, 1992), bifocal display (Spence and Apperley, 1982) or perspective wall (Mackinlay et al., 1991). However, the distortion inherent in these techniques would make it difficult to compare workload data at distant or recurring parts of the task.

Our selected solution uses an interactive overview + detail metaphor, where an overview of the aligned data is shown in one spatial area and details of a selected region are shown in another (Baldonado et al., 2000). This metaphor was chosen because it has been shown to be useful for exploring other temporal data with similar characteristics, e.g., digital video (Casares et al., 2002; Mills et al., 1992) and program execution (Stolte et al., 1999). However, our work extends the basic overview + detail metaphor to support two distinct data sources that are temporally aligned, the pupil data and the model of task execution. Though details and context are not smoothly integrated in this metaphor (Baudisch et al., 2001), we have not yet found this to be a significant limitation for the type of data being analyzed in our work.

Our work is original in that it targets a novel problem domain – analyzing pupillary response in relation to models of task execution. We drew upon and extended existing visualization techniques to produce an effective tool for interactively exploring data in this domain. A small part of our work has been previously presented in (Bailey and Busbey, 2006). This article substantially extends our earlier discussion by including a thorough description of our design goals for the tool, design rationale for the resulting interface, more expansive discussion of the tool's interface components and their use, and empirical results and lessons learned from a user study of the tool.

3. Goals and design evolution

In this section, we describe our design goals for developing a tool that would support interactive analysis of pupillary response data in relation to models of task execution. We then briefly discuss the iterative design process through which our tool was developed, offering design rationale for the particular visualization and interaction design chosen.

3.1. Goals

To generate appropriate design goals for an effective analysis tool, we leveraged our own research experiences, learned from limitations of existing software for analyzing this type of data, and leveraged principles from known visualization techniques. The primary goals were to:

- *Allow pupillary response to be explored relative to a model of task execution and video of on-screen interaction.* To effectively analyze pupillary response, analysts need to navigate and synthesize three distinct sources of data; the raw response data, the abstract model of task execution, and the video of a user's on-screen interaction. There are known techniques for visualizing each of these sepa-

rately; e.g., response data can be graphed over time; task models can be visualized in a tree view; and video can be shown using any media player. However, as none of these data sources make sense independent of the others, the central challenge is to understand how to effectively integrate them within a common visualization aligned to a global timeline. This would allow the aligned data set to be explored from the perspective of any of the data sources.

- *Support rapid exploration of the aligned data set at various levels of detail*. Pupillary response experiments often generate enormous amounts of data, due to extended interaction or high-resolution sampling of the pupil. An analyst thus needs the ability to explore a global view of the aligned data to identify macro-level patterns or to compare patterns at different or recurring parts of the task. At the same time, the analyst also needs the ability to select and examine details in order to assess the response data during execution of specific subtasks.
- *Support a discovery driven analysis process*. When exploring complex relationships in the data set, analysts need to retrieve descriptive statistics for particular parts of the data, record analysis notes, mark points of interest for later review and collaboration, etc. An effective tool should provide user interface controls that support this type of exploratory analysis.
- *Support models of task execution at various levels of detail*. The tasks used in workload experiments can range from basic stimulus-response tasks, e.g., see (Pomplun and Sunkara, 2003), to more complex activities, e.g., see (Schluroff et al., 1986; Verney et al., 2001). An effective tool should thus be able to support models of task execution at various levels of detail, from linear execution sequences to more complex goal hierarchies.
- *Support various data formats and processing algorithms*. Most eye tracking systems generate similar pupil data but in different formats. We wanted our tool to support the specific format of our system, but be easily extensible to others. Since pupil and eye gaze data may be used for myriad purposes, an eye tracking system typically logs all of the data and allows external applications to process the data as needed. Regardless of the format, there are several steps necessary to extract workload from raw pupil data; such as filtering eye blinks, interpolating missed values, calculating relative differences, and smoothing. These techniques are known, but details are often spread throughout the research literature. Our tool could thus serve as a unified resource for analyzing pupillary response, facilitating efficient processing of the data consistent with existing analysis practices. Also, algorithmic extensions to the tool could allow broader and more timely access to newly developed analysis techniques.

Although the current implementation of our tool may not fully meet all of these goals, we felt that it was crucial to define them up front in order to guide later design decisions.

3.2. Audience and user tasks

The audience of the tool is mainly computer scientists and cognitive psychologists working with pupillary response data in relation to task stimuli. This community is currently small, but rapidly growing due to improved hardware at lower costs, better understanding of how to interpret pupil data, and the enduring need to measure mental workload in many user experiments. We also believe that providing more effective analysis tools such as TAPRAV will help enable and encourage more researchers and practitioners to utilize pupillary response in their own work.

A task analysis was performed to identify end user tasks that would benefit most from a new tool in this domain. It was performed by reflecting on our own analysis experiences, data artifacts, research notes, and wish lists generated in prior work (Iqbal et al., 2005; Iqbal and Bailey, 2005) and reviewing procedures, experimental tasks, and analysis diagrams described in related work, e.g., (Hyona et al., 1995; Schluroff et al., 1986; Verney et al., 2001).

Identified tasks included loading and clamping data sources to a common timeline, examining detailed views while maintaining context, retrieving descriptive statistics for specific parts of the task model, comparing workload at different parts of the model, marking points of interest and recording notes, and navigating data sets for multiple users. While not exhaustive, we felt these and related tasks would influence the visualization and interaction design of our tool the most.

3.3. Design evolution

Developing a visualization that was appropriate for our analysis tool presented a significant design challenge. In addition, Amar and Stasko argue that visualizations often suffer worldview gaps, where the specific visualization does not support end users' actual needs (Amar and Stasko, 2004). To address our design challenge and close worldview gaps as best as possible, we felt that it was imperative to develop and test a series of low-fidelity prototypes to more fully explore the design space. Low-fidelity prototypes were developed using paper, sticky notes, colored pens and pencils, etc. (Rettig, 1994) and each prototype was evaluated with 3–5 users, one at a time. Most users had experience with analyzing data from eye tracking experiments or building task models while others had experience with information visualization or user interface design.

As users performed tasks with a prototype, major usability issues were identified by observing parts of the interface that were particularly problematic to comprehend or use, by having users verbalize their ongoing thoughts to determine when expectations did not match the visualization or allowable controls (Rettig, 1994), and by analyzing user feedback about the prototype.

For example, Fig. 1 shows an early prototype in which an overview + detail metaphor was used for the response

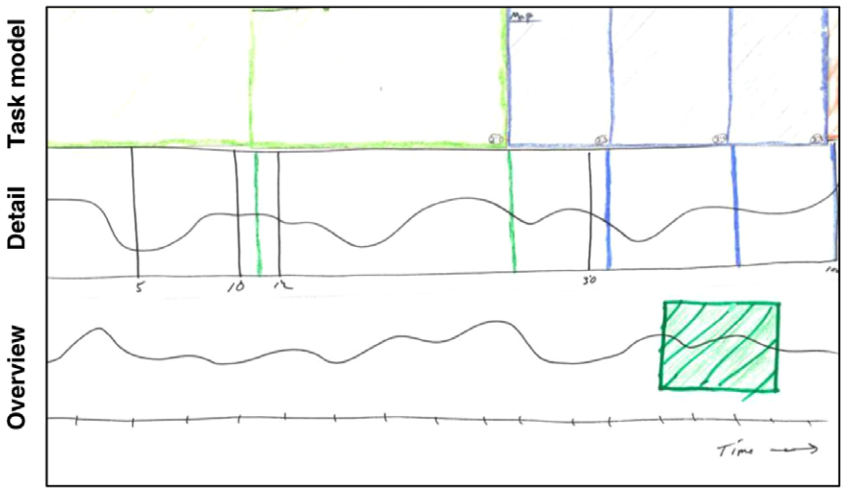

Fig. 1. An early paper prototype of our tool. An overview plus detail model was used for the response graph while a drill-down metaphor was used for exploring the task hierarchy.

graph and a drill-down metaphor was used for exploring the task model. For the graph, the overview frame showed all of the pupil data while the detail frame magnified the selected region. The task model was initially rendered with just the root goal (subtask). Clicking a subtask would split it into its component subtasks and this interaction could be applied recursively, resulting in a visualization similar to Tree-Maps (Johnson and Shneiderman, 1991). However, results from an evaluation showed that this design was too complicated, as users felt the drilling interaction was disorienting during navigation. But, users liked the block metaphor for the subtasks, visual alignment of the data sets, and interactive overview + detail metaphor.

Fig. 2 shows a later paper prototype where the overview + detail metaphor was now applied to both the task model and the response graph. Replacing the drilling interaction, levels of the task model were unfolded into multiple rows and the width of each subtask represented its duration relative to the entire task. The two detail frames were positioned adjacent to each other, aligned at the center of the interface. Magnification lenses were added to both of the overviews, and the data within them was shown within the detail frames, as in (Mills et al., 1992; Stolte et al., 1999).

Feedback from evaluating this prototype was more positive, with users finding it much simpler to understand than prior iterations. However, users felt that the visualization

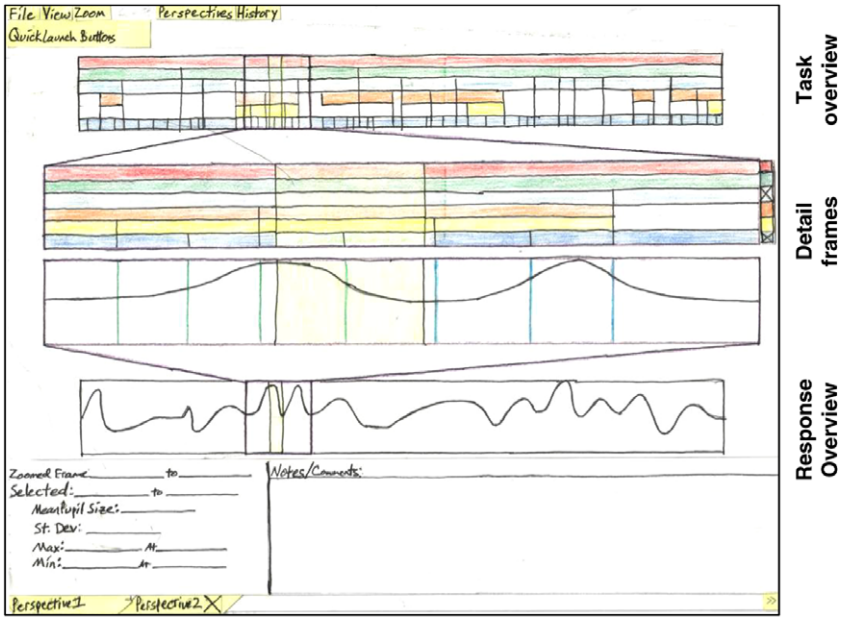

Fig. 2. A later paper prototype of our tool. The overview plus detail metaphor was now extended to both the response graph and the task model, with the two detail frames aligned at the center. A magnification lens (rectangle) in the two overview frames controlled the content in the detail frames.

needed to better emphasize the relationship between the two overview frames and the two detail frames rather than between the two data sources. Our solution was to pair the two overviews and the two detail frames, and place the *pairs* adjacent to each other. This also allowed the two magnification lenses to be collapsed into a single lens, simplifying the overview + detail interaction.

The iterative design process continued until we felt that most of the major usability issues with the visualization and interaction design were resolved. Overall, about five major design iterations were performed, leading to our functional tool TAPRAV.

4. TAPRAV

TAPRAV is a software tool that facilitates rapid, interactive analysis of pupillary response data in relation to hierarchical models of task execution. The primary components of the tool include (i) a visualization of pupillary response aligned to the corresponding model of task execution; useful for exploring relationships between these two data sources; (ii) an interactive overview + detail metaphor, enabling inspection of details while maintaining global context; (iii) synchronized playback of the video of the user's on-screen interaction, allowing better awareness of the current state of the task; and (iv) interaction supporting a discovery driven analysis process; including interactive retrieval of descriptive statistics, marking points of interest, creating views on the data, recording analysis notes, and navigating data sets for multiple users.

4.1. Data background

To offer relevant context for the tool, we review representative data collected in our prior work (Iqbal et al., 2005). In this work, we wanted to analyze how mental workload changes during task execution, e.g., to test whether workload decreases at certain boundaries and whether certain types of subtasks induce more workload than others. As many researchers have speculated that the cost of interruption would be lower during moments of lower mental workload (Miyata and Norman, 1986), results from this analysis would allow us to empirically test this speculation.

In an experiment, users performed several tasks while their pupil size was monitored using an eye tracking system. Tasks included planning routes, editing text documents, and classifying e-mail messages. Fig. 3 shows the route planning task used in the study. Each task lasted about 5 min and the system logged 250 pupil samples per second, along with gaze data. A model for each task was developed and validated using GOMS (Card et al., 1983). The resulting models generally had about 5 levels of subgoal structure and 30–50 low-level operators.

To analyze pupillary response at various parts of the task model, we needed to precisely align the model to the response data, once processed from the raw pupil file. This was achieved by carefully reviewing the cursor and gaze cues in the interaction video to determine the start and end timestamps of each subtask in the model, which could also be used to index the pupillary response data. The task model was then coded and, along with the pupil data, was entered into a numeric spreadsheet.

Performing formal statistical analysis on any part of the data was straightforward. The struggle, however, was to first develop an understanding of how the pupillary response data changed relative to the structure of the task (e.g., to identify patterns and locate unexpected areas of interest) so that we could better direct our analysis efforts, and gain further confidence in and better interpret later results. This required analyzing the pupillary response data at different temporal locations and at different levels of

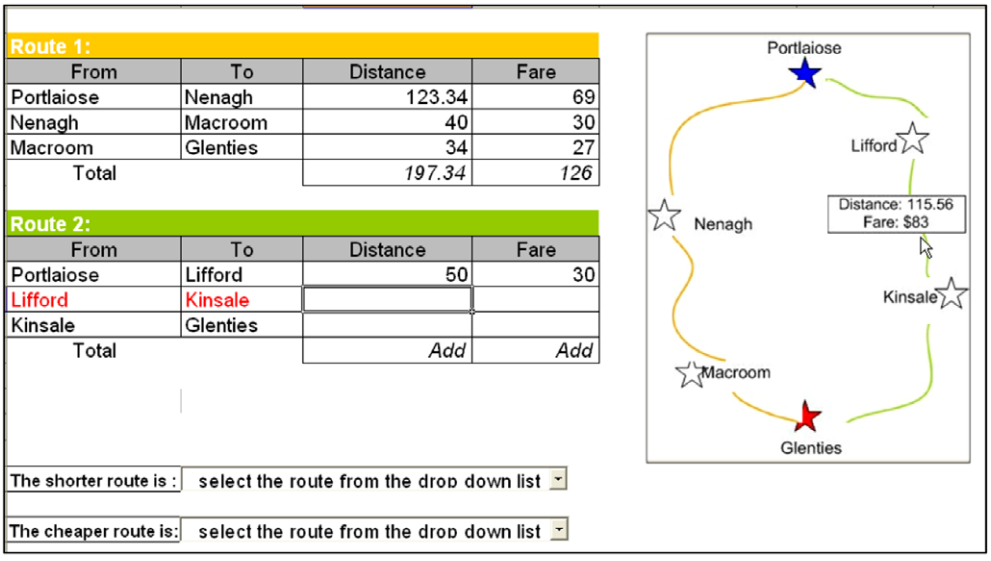

Fig. 3. The route planning task used in our prior work. A user retrieved distance and fare information from the map, entered it into the tables on the left, added the distances and fares, and selected the shorter and the cheaper of the two routes.

hierarchical detail within the task model. Our analysis of the data sets using existing tools required many weeks of laborious effort; as we had to generate, interpret, and compare many independent graphs of various parts of the data and write complex macros to filter and sort the data as needed.

4.2. Visualization

Fig. 4 shows TAPRAV with the task model and pupillary response data from one user performing the route planning task (Fig. 3). Though filtered, the pupil data still contains some noise, but this is not atypical. A few interesting characteristics are immediately visible in the tool. For example, workload (pupillary response) rises rapidly at the onset of the task, fluctuates throughout execution as mental resources are being adaptively allocated, and tails off near the end of the task. These types of observations are consistent with the characteristics that analysts often want to identify or discover, e.g., see (Granholm et al., 1996; Hyona et al., 1995; Schluroff et al., 1986; Verney et al., 2001). Once the visualization and interaction design of the tool are described, we will further elaborate on how the tool can be used to analyze this data (Section 4.8).

The visualization consists of three main components; the pupillary response graph, the model of task execution, and the interactive overview + detail frames.

4.2.1. Response graph

Pupillary response is plotted along the vertical axis over a horizontal timeline. The red line drawn horizontally across the graph represents the baseline value, as workload is computed as the relative increase in pupil size over a baseline. Both the vertical and horizontal axes are of linear scale. Positioning the cursor over a point on the graph displays the response value at that particular time instant, allowing for immediate and detailed data inspection. The cursor can also be quickly moved along the graph in order to follow details of momentary changes in workload.

4.2.2. Task model

A time-based visualization of the task model is provided. A rectangular block represents each subtask in the task model. The width of each block corresponds to its duration relative to the overall task. For example, in Fig. 4, the root node Complete Map Task spans the entire length of the overview frame whereas the lower level node Add Data is much narrower since it lasts for only a few seconds. The name of a task block is drawn within its bounds, space permitting. When the cursor is placed over a block, a tool-tip gives its description and temporal information.

The task model is composed of a collection of blocks and the hierarchy is formed by placing the blocks at each level in the model into successive horizontal rows. The ordering of the blocks along the timeline reflects the ordering of the execution of the subtasks in the model. The model is clamped to the same timeline as the response data, and it is perhaps this property that contributes most to the overall effectiveness of the visualization. The level in the hierarchy, start and end times, and descriptive labels for each subtask block are imported from a task description file. There is one description file for each user performing a task, and it must be defined prior to using our tool.

4.2.3. Overview + detail frames

Given a high sampling rate or lengthy experimental trial, a linear visualization quickly becomes very dense, inhibit-

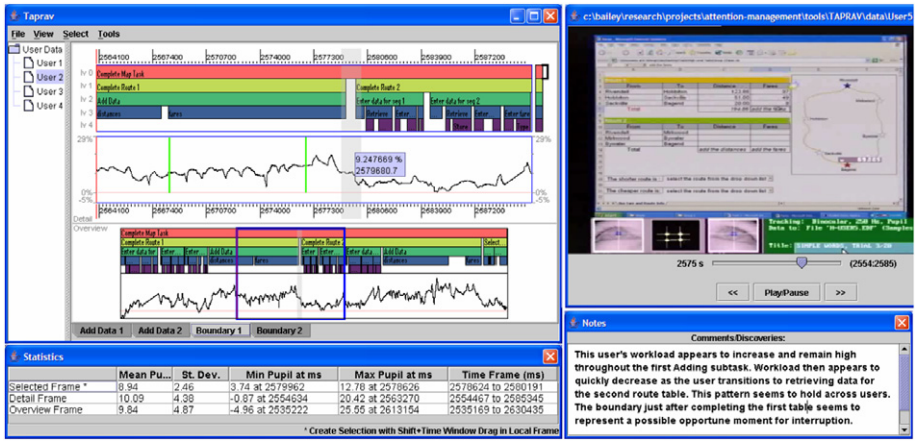

Fig. 4. The main interface screens of TAPRAV. The main window (top left) shows the task model aligned to the pupillary response graph, integrated within an overview + detail metaphor. The two detail and the two overview frames are paired, with the detail frames located above the overview frames. The magnification lens (blue rectangle) allows the analyst to zoom or change the range of data shown in the detail frames. The analyst can click on the timeline to mark points of interest, shown as vertical lines near the middle and left. Multiple views can be created on the data set, which correspond to tabs shown near the bottom of the main window. The statistics window (bottom left) shows the minimum, maximum, and mean values for data within the overview frame, detail frame, and current selection. The video window (top right) shows the user's screen interaction and is synchronized to the response data. The red vertical lines in the detail and overview frames are linked to the video. The Notes window (bottom right) allows the analyst to enter notes about salient features in the current view. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

ing inspection of details. To enable examination of details while maintaining global context, the tool offers an overview + detail metaphor, which was inspired by work on other high-resolution temporal data (Casares et al., 2002; Mills et al., 1992; Stolte et al., 1999).

The overview frame shows the task model aligned to the response graph for the entire data set, which is useful for identifying patterns across different subtasks and locating areas of potential interest. The detail frame shows a specific range of the global data, which is useful for closer inspection of areas of interest.

A magnification lens is shown as a blue rectangle in the overview frame and defines the temporal span of the data shown in the detail frame as well as the temporal span of the video of the user's screen interaction. Analysts can stretch, shrink, and position the lens to temporally zoom the data in the detail frame. A temporal zoom can also be performed from within the detail frame itself. This is done by pressing the mouse button and selecting an area of interest. Once released, the view in the detail frame (and lens) is adjusted to reflect the new selection. This interaction allows analysts to explore the data at various levels of detail while still maintaining global context.

4.3. Video of screen interaction

The video of a user's on-screen interaction can be imported and played in a separate window. The video is aligned to the same timeline as the response data. During playback, a vertical red line is shown moving across the corresponding parts of the response graph and task model in both the overview and detail frames. This allows the analyst to review the current state of the task, better identify a user's location in the task, quickly navigate to locations of interest, and gain confidence when attributing changes in the response data to specific parts of the task.

The start/stop times of the video are defined by the current left/right edges of the magnification lens, allowing the analyst to use the lens to zoom in and out of specific parts of the video. The video window itself has controls for playing, pausing, and scrubbing the video.

4.4. Statistical information

Statistical information can be interactively retrieved for any part of the response graph (and task model). Opened through a menu control, the statistics window shows descriptive statistics for the overview and detail frames, and the currently selected region within the detail frame. Statistics for the overview frame are persistent while statistics for the detail frame are updated in response to the user controlling its temporal range. By pressing a modifier key (⟨shift⟩) while selecting an area of interest, an analyst can retrieve statistics for a range of data within the detail frame itself. This is useful to retrieve statistics for different parts of the data without having to adjust the lens. Holding the modifier disambiguates this particular interaction from the interaction of performing a temporal zoom within the detail frame (see Section 4.2.3).

The descriptive statistics include the mean, standard deviation, minimum, and maximum values, all of which are summarized in a 2D table in the statistics window (see bottom left of Fig. 4). This feature helps the analyst determine if salient features in the visualization may be statistically meaningful and should be subjected to more detailed analysis using existing software packages.

4.5. Points of interest

During the exploration process, the analyst will likely find and want to visually mark notable points of interest on the response graph, as typified in many of the response diagrams shown in (Beatty, 1982; Iqbal et al., 2005; Schluroff et al., 1986; Verney et al., 2001). This allows the analyst to quickly refer to those salient points for later review, comparison, and collaboration.

A marker is a thin vertical line overlaid onto the response graph. To insert a marker, the analyst clicks at the desired point on the timeline in the detail frame. Any number of markers can be added and whether markers are visible can be toggled with a menu control.

4.6. Multiple views and notes

TAPRAV allows analysts to save specific views of the data and return to them later. A new view tab is created through a menu control and, when created, the tool records the position and size of the magnification lens, visual markers, recorded notes, and current selection in the detail frame. Views are navigated via a tabbed interface, which is shown at the bottom of the visualization panel. This feature allows analysts to save snapshots of interesting parts of the data while continuing to explore other parts in the same session. For each view, the analyst can enter comments into a Notes dialog and these are available whenever that particular view is active.

4.7. Data for multiple users

A data set (response data, task model, and video) can be imported for any number of users. User data sets can be accessed through a tree view on the left panel and new ones can be added using a menu control. Once a data set is imported, all the interactions previously described are available. Our ongoing work is investigating how to visualize data sets that have been aggregated across users, which is challenging because the durations of subtasks are almost always different.

The current state of the analysis can be saved into a session file. Loading the session file into the tool at a later time enables the analyst to quickly continue from where they last left off.

4.8. Putting it all together

We now elaborate on how TAPRAV can be used to explore the data discussed in Section 4.1. After launching the tool, the task model and pupil data are imported via menu controls. The tool filters eye blinks and saccades, keeping just the fixations, and linearly interpolates any missing values (which appear as 0's in the pupil data file). It then retrieves the baseline value, guided by user input, and computes relative differences at each sample point.

Fig. 5. (a) The initial state of the tool after data sets for multiple users were imported, each consisting of the pupil data, task model, and video of screen interaction. The dashed rectangles were superimposed to show where pupillary response is noticeably higher for this particular user. The lens initially spans the entire overview frame, thus the detail frame matches the overview. (b) The lens has been adjusted to cover the first adding subtask. The average response value within the detail frame is 12.2%, indicated in the second row of the statistics window. (c) The lens has been adjusted to cover the second adding subtask, and the average response value within the detail frame is just over 14%. (d) The lens has been adjusted to span an area slightly larger than the boundary between the first and second routes. The boundary region has been selected within the detail frame (shown in grey), and the relevant values are retrieved from the first row of the statistics window. (e) The lens has been adjusted to span an area around the boundary between completing the second route and making the selections, and the specific boundary region has been selected within the detail frame.

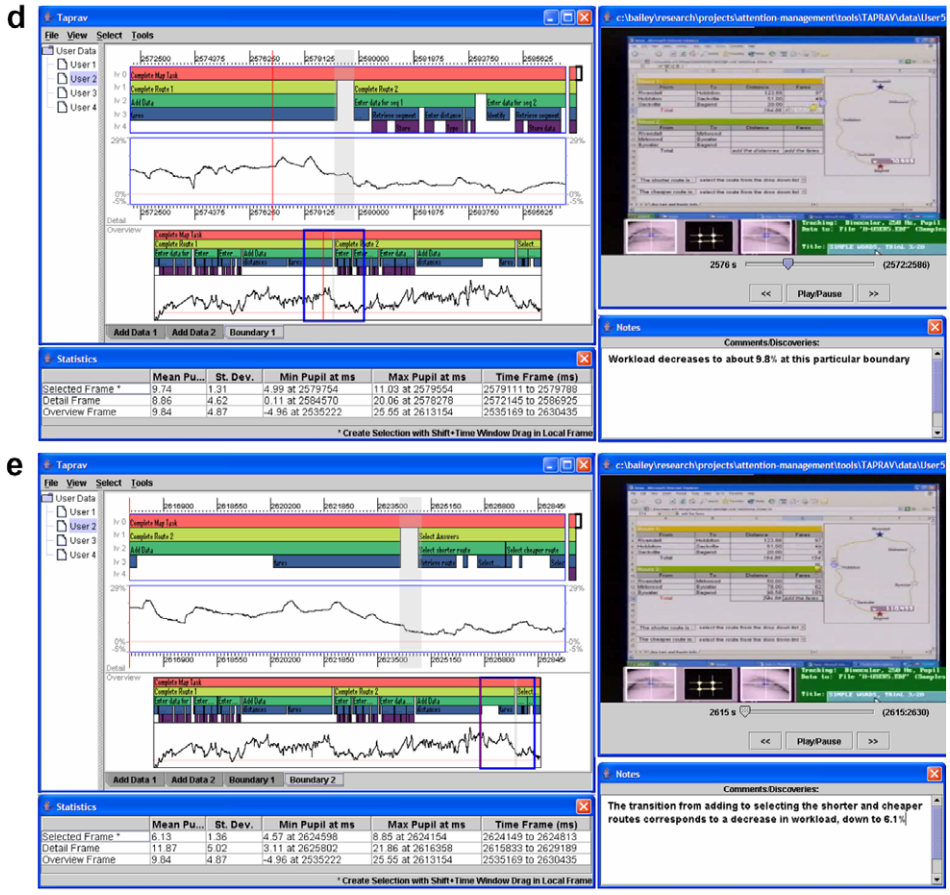

Fig. 5 (*continued*)

From a menu control, the view is trimmed to show only the pupil data within range of the task's execution. The video is imported and an offset is entered synchronizing it to the global timeline. This process is repeated to import data sets for multiple users and results in the basic visualization shown in Fig. 5a.

For this analysis, we wanted to know which parts of the task had higher and lower workload and whether workload decreased at any particular subtask boundaries. A boundary is the short period between any two adjacent subtasks. From the detail frame, which currently shows the entire data set as the lens initially spans the entire overview frame, we observe that the two adding subtasks appear to have higher workload (highlighted in Fig. 5a). To explore this further, we position the lens around the first Add Data subtask and the statistics window shows that the average value of the pupil data within the detail frame is about 12.2% over the baseline (see Fig. 5b).

To allow later review, the current view tab is named (Add Data 1) and a new view tab is created (Add Data 2). The lens is positioned over the second Add Data subtask and the average response is 14% (see Fig. 5c), meaningfully higher than the first adding subtask. Reviewing the video of screen interaction at both subtasks, which can be done quickly by navigating the view tabs, offers a plausible explanation; the second adding subtask required more complex calculations. This explanation is entered into the Notes window. We also find that workload is lower when the user entered data into the first (6.98%) and second (6.73%) tables and when selecting the shorter and cheaper routes (5.84%).

Further inspection of the overview frame shows that the response data noticeably decreases at two boundaries; between Complete Route 1 and Complete Route 2, and between Complete Route 2 and Select Answers. Following a similar process (see Figs. 5d and e), we find that the average response during each boundary (9.74% and 6.13%, respectively) is lower than the preceding subtask. Note that in Figs. 5d and e, statistics are retrieved by sweeping an area within the detail frame itself, allowing statistical data to be retrieved without having to adjust the magnification lens. By switching among the data sets for the other users, we are able to quickly determine that this pattern is fairly consistent. These boundaries thus seem to represent cognitive breakpoints in the task and may represent lower cost moments for interruption (Miyata and Norman, 1986), which we were able to empirically confirm in a later experiment (Iqbal and Bailey, 2005).

Realizing this scenario with existing tools would be cumbersome since the tools either do not support a

visualization of the aligned data sets or the resulting visualization is not interactive. In contrast, TAPRAV allows this scenario to unfold seamlessly, as the analysis process is directly supported by the visualization and interaction design of the tool. This meaningfully reduces the effort required to explore relationships among the pupil data, task model, and interaction video.

The use of TAPRAV was described only for our data sets and analysis needs, but it could also be leveraged to conduct exploratory analyses of other similar data sets, such as those reported in (Beatty, 1982; Granholm et al., 1996; Hyona et al., 1995; Schluroff et al., 1986).

4.9. Implementation

TAPRAV was written in Java and consists of about 5,000 lines of code. Java was chosen because it executes on multiple platforms and has numerous external APIs available. The task model and response graph are drawn using Java2D, the interface controls use Java Swing, and video is handled using QuickTime for Java. The tool has been tested on both Mac and PC machines.

The task model is parsed from a description file that must be defined by the analyst using the notation expected by our tool. The notation consists of a single ⟨task⟩ tag that can be arbitrarily nested and sequenced to define the hierarchical and sequential structure of a task's execution. For example, Fig. 6 shows part of the file describing one user's execution of the route planning task. The level of detail specified in the task description is determined by the analyst based on the types of research questions being posed.

The ⟨task⟩ tag supports attributes for specifying the start and end timestamp of the containing subtask and these timestamps index the pupillary response data. To determine these values, the cursor and gaze cues in the interaction video must be carefully analyzed to determine precisely when each subtask starts and ends. This analysis must be performed prior to using our tool.

```
<!-- User task model; video offset = 643250 -->
<task name = "Complete Map Task" offset = "2506686">
    <task name = "Complete Route 1">
        <task name = "Enter data for segment 1">
            <task name = "Identify segment">
                <start>28483</start>
                <end>29687</end>
            </task>
            <task name = "Retrieve segment">
                <task name = "Locate segment in map">
                    <start>29845</start>
                    <end>30629</end>
                </task>
                <task name = "Store data">
                    <start>30849</start>
                    <end>33389</end>
                </task>
            </task>
            ...
        </task>
    </task>
</task>
```

Fig. 6. Part of a task description file for one user's execution of the route planning task.

Because users typically perform tasks at different speeds, a separate description file needs to be created for each user/task combination. Similar description files would be created for other experimental tasks, e.g., to indicate phases of sentence comprehension (Schluroff et al., 1986).

The pupillary response data is parsed from a known data file containing time-stamped samples of pupil size, which is generated by the eye tracking system. To enable video synchronization, the eye tracker records a start event into its data file and overlays the value of its system clock onto the video. Once the frame of video showing the timestamp of the start symbol is located, the offset (current media time of the video) is entered into the tool, synchronizing the video to the global timeline.

5. User study

A user study was conducted to assess how well users could utilize TAPRAV to explore and understand relationships between pupillary response and a model of task execution for an ecological data set, study how users interact with the tool during the analysis process, gain high-level reactions about different parts of the interface, and identify additional usability issues. We felt that assessing and improving the interaction design of the tool through a controlled study was an important and necessary first step in preparing to later study the use of the tool in the field.

5.1. Users and tasks

Eight users (1 female) participated in the study and the average age was 25 (SD = 2.8). All users were either undergraduate or graduate students in Computer Science at our institution. Most users had at least some experience analyzing large data sets using various visualization tools, but none had specific expertise in this particular task domain. However, we did not believe that this would be a significant limitation, as our focus was on assessing the basic utility of the tool.

User tasks consisted of interacting with the tool to answer a given set of analysis questions pertaining to an existing data set (pupil data, task model, and interaction video), which was similar to the data set described in Sections 4.1 and 4.8. Three types of questions were posed; (D) *directed questions*, where users determined the average pupillary response (workload) during specific subtasks or boundary regions within the task model; (E) *exploratory questions*, where users compared workload at various parts of the task to determine where workload was lower/higher and to offer a plausible explanation; and (O) *open-ended questions*, where users freely explored the data set and identified any interesting relationships or patterns. Users could use either the Notes window in the tool or a supplied paper

notebook to record their solutions, explanations, and observations. The specific analysis questions were:

- (D) What was the user's average workload when adding fare information in the first table?
- (D) What was the average workload when adding distance information in the second table?
- (E) Which subtask at Level 2 had the lowest workload? Can you explain why?
- (E) Which subtask at Level 2 had the highest workload? Can you explain why?
- (D and E) What was the average workload at the boundary between finishing the second table and starting the selection of the shorter route? Is this value less than the preceding subtask?
- (O) Are there any other interesting patterns of workload that you see in the data set?

These questions were based on those asked in our prior work (Iqbal et al., 2005), and are representative of the types of questions asked in other related work, e.g., (Backs and Walrath, 1992; Granholm et al., 1996; Schluroff et al., 1986; Verney et al., 2004). The questions were carefully constructed to prompt the user to interact with each of the major interface components of the tool, but without needing to specifically instruct users to do so.

5.2. Procedure and measurements

Upon arriving at the laboratory, we went through an informed consent process with the user. The user was then seated at a desk with a standard desktop computer running TAPRAV. The data set was already imported into the tool. We provided a brief overview of the various interface components and gave an explanation of the data set. The user was allowed to practice using the tool and ask any questions. The user then began interacting with the tool to answer the analysis questions, which were provided on a sheet of paper. The user's screen interaction was recorded using Camtasia, a commercial software tool that captures the video frame buffer.

Once finished with using the tool, the user filled out a paper questionnaire. The user was asked to rate the usefulness of the main components of the interface; including the aligned visualization, overview + detail metaphor, synchronized video window, ability to retrieve statistical data, and ability to create multiple views on the data. Responses were structured using a 7-point Likert scale, ranging from Not Useful (1) to Very Useful (7). Users were also asked to explain their reasoning for each rating. Finally, users were asked to describe any particular strengths or weaknesses of the tool. The entire experimental session lasted less than 60 min.

Measurements included the correctness of users' answers to the analysis questions, observations of how the tool was utilized, and users' ratings and responses on the post-task questionnaire.

5.3. Results

All of the users were able to derive answers for the analysis questions in the allotted time. Users interacted with the tool in a common and expected pattern. They would use the overview frame to determine the next target subtask or boundary area, size and position the magnification lens around an area slightly larger than the target area, move the cursor to the detail frame, select a more specific region and record the statistical values, review the video (if necessary), and repeat. TAPRAV thus seemed to facilitate a natural problem solving strategy for analyzing the data.

As indicated in Table 1, nearly all of the users were able to produce correct solutions. If an error was made, it was usually because the user did not notice or forgot that each subtask block was associated with a specific level in the task model. This property needs to be made more salient in the visualization, e.g., by highlighting the text label and border for a level whenever the cursor enters any of its subtask blocks. For the open-ended question, users generally responded with which types of subtasks had the highest (e.g., adding) or lowest (e.g., selection of routes) workload. One user also noticed that workload seemed to momentarily decrease at boundaries higher in the task model; a result that we reported in prior work, but that required several weeks of laborious analysis effort to determine. Overall, these results show that users, who have limited or no experience in this particular domain, are able to successfully interact with our tool to answer realistic analysis questions and discover meaningful relationships in the data.

Users rated the overview + detail model ($\mu = 6.4$, SD = 1.1), aligned visualization ($\mu = 6.0$, SD = 1.4), and ability to retrieve descriptive statistics ($\mu = 5.7$, SD = 1.6)

Table 1
User responses to analysis questions compared to correct solutions

Question	(Mean, SD) or number correct	Correct solution
Average workload when adding fare information in the first table	(11.8, 1.04)	12.1
Average workload when adding distance information in the second table	(13.3, 1.44)	13.7
Workload at the boundary between finishing the second table and selecting shorter route	(6.7, 1.25)	6.6
L2 subtask with lowest workload	5/8	Entering data for the third segment in the first table
L2 subtask with highest workload	6/8	Adding distance and fare data in the second table
Is the workload at the boundary (from previous question) less than the workload during its preceding subtask?	8/8	Yes

as the more useful parts of the tool; while rating the ability to create multiple views ($\mu = 4.9$, SD = 1.9) and view the video of the user's interaction ($\mu = 4.5$, SD = 2.3) as the less useful parts. This pattern was likely due to users needing to use the first three components more than the latter two in the study. User feedback highlighted the utility of using different components of the interface to explore relationships in the data:

> "Having the task model aligned with the data allowed an easy way to match changes in the pupil size to the tasks in which they occurred."

> "It was useful that the program automatically calculated statistics based on the user selection. This was fast, convenient and easy to learn/use."

> "The pupil data and task model aligned together was useful because it tags the data with semantic information that are too hard (or impossible) to extract from the raw data."

> "It was very useful to see detailed information. What made it more useful was to see it in the overall context of the entire data set. This gave the user a better grounding of the detail model with respect to the overview."

> "[The synchronized video] allowed me to go back and see specifically what the user was doing. For example, I was able to see the data the user saw without having to ask for it or look it up."

Several lessons were learned about how to improve the interaction design of the tool. First, analysis notes entered into the Notes dialog should be globally available. Currently, notes shown within the dialog are associated with the current view and change each time a different view is selected. However, users wanted the notes to be linked directly to the data, not to the current view, and explained that this was mostly why they preferred the use of the paper notebook. For example, one user stated, "I did not like that my notes went away when I switched view tabs" while another user stated "...while working and comparing data sets I expected my notes to 'travel' with me. Each view seemed to have its own notepad..." Among several alternative designs, our tool now allows note icons to be inserted within the detail frame via a menu control and the icons initially appear at the current position of the time marker. Users are then able to enter the associated text and later review it by positioning the cursor over the icon, causing a tool tip to appear. This re-design has replaced the current Notes window.

Second, the user should not have to press and hold the modifier key while selecting an area of interest within the detail frame to retrieve statistics. When an area within the detail frame is selected without the modifier key, the interface zooms to that area. However, users almost never used this feature; rather, they almost always used the magnification lens to adjust the area of detail and were often confused when the interface jumped to another range of data. As a result, we modified our tool such that the more common case of selecting an area of interest to retrieve statistics no longer requires the use of a modifier key, while zooming into the data does.

Third, the summary statistics for the selected area within the detail frame should be displayed with a integrated callout or similar technique. Users disliked the current display configuration, as they had to repeatedly switch their visual attention from the area of interest to the statistics window. For example, one user commented, "it was useful to know the statistical numbers, but it was sometimes difficult to contextualize them." Several users noted that the statistical data would be easier to attend to if it were placed closer to the selected area within the detail frame.

Finally, users should be able to select multiple areas of interest within the detail frame. The concept of creating multiple views was generally liked, as this would allow workload at different parts of the task to be compared. The problem was that only one view (and summary statistics) could be seen at any given time. Consistent with suggestions from several users, our future work seeks to address this and the previous issue by allowing multiple selections within the detail frame to be active at once and showing a visual callout with statistical information for each one.

6. Discussion and future work

We describe how our current implementation has heretofore met our project design goals. To allow pupillary response to be explored in relation to a task model, we developed an interactive visualization that shows the model temporally aligned to the response data. The aligned data set was integrated within an interactive overview + detail metaphor. This allows the data to be examined at various levels of detail while maintaining context of the overview data. To facilitate discovery driven analysis, the tool includes interaction for marking points of interest, creating views, recording notes, loading data sets for multiple users, and saving/loading analysis sessions.

The tool offers an XML-based notation for describing models of task execution at various levels of detail, from flat execution sequences to those with hierarchical goal structures. Start and end times are assigned to each subtask, allowing the task's execution structure to be clamped to the response data. Timestamps are determined independent of using our tool at the present time, but our future work seeks to allow interactive construction of the task models and specification of the start/end timestamps for the subtasks. Though relatively simple, our notation is currently sufficient for describing most of the tasks that have been reported for experiments involving pupillary response, e.g., see tasks used in (Backs and Walrath, 1992; Iqbal et al., 2004, 2005; Schluroff et al., 1986; Verney et al., 2001, 2004).

In terms of pupil data formats, our tool currently parses the data format for a specific, but commonly used eye tracking system (Eye Link II), while the architecture of our tool allows for plug-ins to parse different file formats and load the data into a common data structure. To extract workload from the raw pupil and gaze data, our tool implements known algorithms for filtering eye blinks, interpolating missed values, applying various smoothing filters, and computing increases relative to a given baseline or deriving it from a given time span. By integrating these algorithms, however, our tool can help facilitate the use of accepted practices for analyzing this type of data. Overall, we have made significant strides towards meeting our design goals.

Interpreting pupillary response data can be challenging, as there is no commonly accepted scale of mental workload. Recent research shows promise for addressing this issue. For example, Marshall has recently proposed the Index of Cognitive Activity (ICA), which is computed from pupillary response (Marshall, 2002). An exciting possibility is that our tool could use the ICA as the scale for the *Y*-axis and automatically map the raw response data to this scale. This would facilitate common interpretation of the data and enable analysts to utilize this new metric without needing to know its algorithmic details.

Our tool currently supports pupillary response as the measure of workload. However, many other physiological measures such as electroencephalogram (Schacter, 1977), heart rate variability (Rowe et al., 1998), and event-related potential (Kok, 1997) can also provide an effective measure of workload. If these measures could be mapped onto the ICA or other common scale, then our tool could be extended to support them. Otherwise, the visualization and interaction demonstrated in our tool could serve as a template for designing similar tools for these measures.

Our visualization uses a single overview and detail frame. While this has proven sufficient for our data, other tools for exploring high-resolution temporal data (e.g., video) allow users to create a hierarchy of detail frames, where each subsequent frame shows successively narrower ranges of the data (Casares et al., 2002; Mills et al., 1992; Stolte et al., 1999). If necessary, this interface feature could be integrated into future iterations of our tool.

Though we presented results from a user study, we have not yet conducted a formal comparison between the use of our tool and existing methods of exploring pupillary response in relation to models of task execution. However, as part of our ongoing research in this area, we have gained considerable experience in using both existing tools and TAPRAV to analyze dozens of data sets. Based on this experience, we believe that the use of our tool can indeed reduce the effort required to analyze this type of data and communicate results among the research team; but the differences need to be quantified in an empirical study. It is also important to point out that the use of our tool is expected to complement existing practices for analyzing this type of data, not replace them; and its use should be most beneficial during the exploratory stages of data analysis.

Beyond fixing known usability issues, we have several directions for future work. First, we want to conduct field studies to better understand how well our interactive tool supports analysis of pupillary response relative to task execution and how its use fits into existing work practices. Second, we want to include a direct manipulation interface for constructing the task models, as the current method requires them to be defined offline. In addition, we would like to explore the value of extending the description language and corresponding visualization to support more complex execution sequences. Finally, we want to implement additional algorithms for smoothing and analyzing the response data as well as for aggregating data sets across users.

7. Conclusion

As a reliable indicator of mental workload, pupillary response is being increasingly leveraged as a research instrument, e.g., to identify lower cost moments for interruption, evaluate complex interfaces, and understand psychological processes. Yet existing software tools are not sufficient for analyzing pupillary response data, as it typically needs to be explored in relation to a model of the corresponding task's execution. To address this emerging need, we have developed a new interactive analysis tool, TAPRAV. The tool demonstrates an effective technique for visualizing a hierarchical model of task execution aligned to a continuous measure of workload, integrated within an overview + detail metaphor. This visualization was fully implemented in the tool along with user interface controls supporting discovery driven analysis. TAPRAV is publicly available at: http://orchid.cs.uiuc.edu/projects/TAPRAV. The use of this tool can considerably reduce the effort needed to analyze pupillary response in relation to models of task execution, which may further enable and encourage the use of this type of data in both research and practice.

Acknowledgement

This work was supported in part by a grant from the National Science Foundation under award no. IIS 05-34462.

References

Altmann, E.M., Trafton, J.G., 2004. Task interruption: resumption lag and the role of cues. In: Proceedings of the 26th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates.

Amar, R., Stasko, J., 2004. A knowledge task-based framework for design and evaluation of information visualizations. In: Proceedings of the IEEE Symposium on Information Visualization. IEEE Computer Society, pp. 143–150.

Backs, R.W., Walrath, L.C., 1992. Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. Applied Ergonomics 23, 243–254.

Bailey, B.P., Busbey, C.W., 2006. TAPRAV: a tool for exploring workload aligned to task models. In: Proceedings of the International Conference on Advanced Visual Interfaces (AVI). ACM Press, pp. 467–470.

Baldonado, M.W., Woodruff, A., Kuchinsky, A., 2000. Guidelines for using multiple views in information visualization. In: Proceedings of the International Conference on Advanced Visual Interfaces (AVI). ACM Press, pp. 110–119.

Baudisch, P., Good, N., Stewart, P., 2001. Focus plus context screens: combining display technology with visualization techniques. In: Proceedings of the ACM Symposium on User Interface Software and Technology. ACM Press, pp. 31–40.

Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin 91 (2), 276–292.

Card, S., Moran, T., Newell, A., 1983. The Psychology of Human–Computer Interaction. Lawrence Erlbaum Associates, Hillsdale, NJ.

Casares, J., Long, A.C., Myers, B., Stevens, S., Corbett, A., 2002. Simplifying video editing with silver. In: Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM Press, pp. 672–673.

Granholm, E., Asarnow, R.F., Sarkin, A.J., Dykes, K.L., 1996. Pupillary responses index cognitive resource limitations. Psychophysiology 33 (4), 457–461.

Hart, S.G., Staveland, L.E., 1988. Development of a multi-dimensional workload rating scale: results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. Elsevier, Amsterdam, The Netherlands, pp. 138–183.

Hornbæk, K., Bederson, B.B., Plaisant, C., 2002. Navigation patterns and usability of zoomable user interfaces with and without an overview. ACM Transactions on Computer–Human Interaction 9 (4), 362–389.

Hyona, J., Tommola, J., Alaja, A., 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. The Quarterly Journal of Experimental Psychology 48A (3), 598–612.

Iqbal, S.T., Bailey, B.P., 2005. Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In: Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM Press, pp. 1489–1492.

Iqbal, S.T., Zheng, X.S., Bailey, B.P., 2004. Task evoked pupillary response to mental workload in human–computer interaction. In: Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM Press, pp. 1477–1480.

Iqbal, S.T., Adamczyk, P.D., Zheng, S., Bailey, B.P., 2005. Towards an index of opportunity: understanding changes in mental workload during task execution. In: Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM Press, pp. 311–320.

Ivory, M.Y., Hearst, M.A., 2001. The state of the art in automating usability evaluation of user interfaces. ACM Computing Surveys 33 (4), 470–516.

John, B.E., 1995. Why GOMS? Interactions 2, 80–89.

John, B.E., Kieras, D.E., 1996. The GOMS family of user interface analysis techniques: comparison and contrast. ACM Transactions on Computer–Human Interaction 3 (4), 320–351.

John, B.E., Vera, A., Matessa, M., Freed, M., Remington, R., 2002. Automating CPM-GOMS. In: Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM Press, pp. 147–154.

Johnson, B., Shneiderman, B., 1991. Tree-Maps: a space-filling approach to the visualization of hierarchical information structures. In: Proceedings of the IEEE Conference on Information Visualization. IEEE Computer Society, pp. 284–291.

Just, M.A., Carpenter, P.A., Miyake, A., 2003. Neuroindices of cognitive workload: neuroimaging, pupillometric, and event-related potential studies of brain work. Theoretical Issues in Ergonomics 4, 56–88.

Kirwan, B., Ainsworth, L.K., 1992. A Guide to Task Analysis. Taylor & Francis, Ltd.

Kok, A., 1997. Event-related-potential (ERP) reflections of mental resources: a review and synthesis. Biological Psychology 45, 19–56.

Kramer, A.F., 1991. Physiological metrics of mental workload: a review of recent progress. In: Damos, D.L. (Ed.), Multiple-Task Performance. Taylor and Francis, London, pp. 279–328.

Mackinlay, J.D., Robertson, G.G., Card, S.K., 1991. The perspective wall: detail and context smoothly integrated. In: Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM Press, pp. 173–176.

Marshall, S.P., 2002. The index of cognitive activity: measuring cognitive workload. In: Proceedings of the 7th Conference on Human Factors and Power Plants. IEEE Computer Society, pp. 7.5–7.9.

Marshall, S.P., 2003. New techniques for evaluating innovative interfaces with eye tracking. In: Proceedings of the ACM Symposium on User Interface Software and Technology, Keynote Talk.

Mills, M., Cohan, J., Wong, Y.Y., 1992. A magnifier tool for video data. In: Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM Press, pp. 93–98.

Miyata, Y., Norman, D.A., 1986. Psychological issues in support of multiple activities. In: Norman, D.A., Draper, S.W. (Eds.), User Centered System Design: New Perspectives on Human–Computer Interaction. Lawrence Erlbaum, Hillsdale, NJ, pp. 265–284.

Moray, N., Johansson, J., Pew, R., Rasmussen, J., Sanders, A.F., Wickens, C.D., 1979. Mental Workload, It's Theory and Measurement. Plenum Press, New York.

Oviatt, S., Coulston, R., Lunsford, R., 2004. When do we interact multimodally? Cognitive load and multimodal communication patterns. In: Proceedings of the Sixth International Conference on Multimodal Interfaces. ACM Press, pp. 129–136.

Paganelli, L., Paternò, F., 2003. Tools for remote usability evaluation of web applications through browser logs and task models. Behavior Research Methods, Instruments, and Computers 35 (3), 369–378.

Paternò, F., Mancini, C., Meniconi, S., 1997. ConcurTaskTrees: a diagrammatic notation for specifying task models. In: Proceedings of the IFIP TC13 International Conference on Human–Computer Interaction. Chapman and Hall, pp. 362–369.

Pomplun, M., Sunkara, S., 2003. Pupil dilation as an indicator of cognitive workload in human–computer interaction. In: Proceedings of the 10th International Conference on Human–Computer Interaction. Lawrence Erlbaum Associates, pp. 542–546.

Rettig, M., 1994. Prototyping for tiny fingers. Communications of the ACM 37 (4), 21–27.

Rowe, D.W., Sibert, J., Irwin, D., 1998. Heart rate variability: indicator of user state as an aid to human-computer interaction. In: Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM Press, pp. 480–487.

Sarkar, M., Brown, M.H., 1992. Graphical fisheye views of graphs. In: Proceedings of the ACM Conference on Human Factors in Computing Systems. ACM Press, pp. 83–91.

Schacter, D.L., 1977. Eeg theta waves and psychological phenomena: a review and analysis. Biological Psychology 5 (1), 47–82.

Schluroff, M., Zimmermann, T.E., Freeman, R.B., Hofmeister, K., Lorscheid, T., Weber, A., 1986. Pupillary responses to syntactic ambiguity of sentences. Brain and Language 27, 322–344.

Spence, R., Apperley, M., 1982. Database navigation: an office environment for the professional. Behavior and Information Technology 1 (1), 43–54.

Stolte, C., Bosch, R., Hanrahan, P., Rosenblum, M., 1999. Visualizing application behavior on superscalar processors. In: Proceedings of the IEEE Symposium on Information Visualization. IEEE Computer Society, pp. 10–17.

Takahashi, K., Nakayama, M., Shimizu, Y., 2000. The response of eye-movement and pupil size to audio instruction while viewing a moving target. In: Proceedings of the ACM Conference on Eye Tracking Research and Applications. ACM Press, pp. 131–138.

Verney, S.P., Granholm, E., Dionisio, D.P., 2001. Pupillary responses and processing resources on the visual backward masking task. Psychophysiology 38 (1), 76–83.

Verney, S.P., Granolm, E., Marshall, S., 2004. Pupillary responses during the visual backward masking task predict cognitive ability. International Journal of Psychophysiology 52, 23–36.

Wickens, C.D., 2002. Multiple resources and performance prediction. Theoretical Issues in Ergonomic Science 3 (2), 159–177.

Zacks, J.M., Tversky, B., 2001. Event structure in perception and conception. Psychological Bulletin 127, 3–21.