Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data

BEN STEICHEN, CRISTINA CONATI, and GIUSEPPE CARENINI,
University of British Columbia

Information visualization systems have traditionally followed a one-size-fits-all model, typically ignoring an individual user's needs, abilities, and preferences. However, recent research has indicated that visualization performance could be improved by adapting aspects of the visualization to the individual user. To this end, this article presents research aimed at supporting the design of novel user-adaptive visualization systems. In particular, we discuss results on using information on user eye gaze patterns while interacting with a given visualization to predict properties of the user's visualization task; the user's performance (in terms of predicted task completion time); and the user's individual cognitive abilities, such as perceptual speed, visual working memory, and verbal working memory. We provide a detailed analysis of different eye gaze feature sets, as well as over-time accuracies. We show that these predictions are significantly better than a baseline classifier even during the early stages of visualization usage. These findings are then discussed with a view to designing visualization systems that can adapt to the individual user in real time.

Categories and Subject Descriptors: H.5.m [**Miscellaneous**]

General Terms: Human Factors, Experimentation

Additional Key Words and Phrases: Adaptive information visualization, eye tracking, adaptation, machine learning

ACM Reference Format:
Ben Steichen, Cristina Conati, and Giuseppe Carenini. 2014. Inferring visualization task properties, user performance, and user cognitive abilities from eye gaze data. ACM Trans. Interact. Intell. Syst. 4, 2, Article 11 (July 2014), 29 pages.
DOI: http://dx.doi.org/10.1145/2633043

1. INTRODUCTION

Information visualization is a thriving area of human-computer interaction that aims to help users in managing and understanding increasing amounts of information. Although visualization systems have gained in terms of general usage and usability, they have traditionally been designed using a one-size-fits-all approach, typically ignoring an individual user's needs, abilities, and preferences. To better assist each user during visualization tasks, recent research has started to investigate novel user-adaptive visualizations that can dynamically infer relevant user characteristics and provide appropriate interventions tailored to these characteristics. Initial research on user-adaptive visualizations has already provided evidence for improved user performance (e.g., time on task, task accuracy), such as by using click behavior to infer and

The reviewing of this article was managed by associate editor Judy Kay.

Authors' addresses: B. Steichen, C. Conati, and G. Carenini, Department of Computer Science, University of British Columbia, Department of Computer Science, Vancouver, Canada; emails: steichen@cs.ubc.ca, conati@cs.ubc.ca, carenini@cs.ubc.ca.

DOI: http://dx.doi.org/10.1145/2633043

adapt to suboptimal usage patterns [Gotz and Wen 2009], or by considering a user's visualization selections to infer and adapt to a user's visualization expertise and preferences [Grawemeyer 2006]. In terms of intervention mechanisms, these initial systems have typically investigated recommending visualizations that are most suitable for the current task and/or appropriate for a particular user's preference and expertise.

Our long-term goal is to extend this research on user-adaptive visualization in a number of aspects. First, we aim to expand the set of features that the system can adapt to by including visualization task properties (e.g., task type, task difficulty), as well as a user's properties beyond expertise and performance, such as cognitive abilities that have been shown to influence visualization performance. Second, although existing research has looked at improving visualization performance solely using information on a user's direct interaction (e.g., mouse clicks), we aim to provide assistance exploiting additional, potentially complementary data sources (e.g., eye tracking). Third, whereas existing work has focused on interventions that recommend alternative visualizations, we envision to also deliver interventions that can dynamically help the user with the current visualization (e.g., through highlighting relevant visualization elements).

In this article, we address the first two aspects by investigating to what extent a variety of visualization task properties (task type, complexity, and difficulty), the user's performance (in terms of task completion time), the user's visualization expertise, and three different cognitive abilities (perceptual speed, visual working memory, and verbal working memory) can be inferred from a user's eye gaze behavior. For all of these dimensions, we found statistically significant results, except for user visualization expertise.

We focus on gaze behavior because visual scanning and processing are fundamental components of working with any visualization (and the only components for noninteractive visualizations). Specifically, we ask the following two research questions:

Q1. To what extent can a user's current task, performance, and / or long-term cognitive abilities and visualization expertise be inferred from eye gaze data?
Q2. Which gaze features are the most informative?

The motivation of this work is twofold. First, to provide appropriate support, an adaptive system needs to know about the user's current task characteristics, as well as her expected performance. For example, if the system knows that the user is currently performing a "filter" task (i.e., trying to find data cases that satisfy a particular condition [Amar et al. 2005]) and appears to be slow (i.e., the user is predicted to have a high completion time), the system could adaptively de-emphasize nonrelevant data to reduce the user's cognitive load. Correspondingly, if a system knows individual *user* characteristics, it will be able to provide user-specific support. For example, since certain cognitive ability levels have already been shown to lead to lower performance (e.g., low perceptual speed leads to decreases in speed and accuracy [Conati and Maclaren 2008; Velez et al. 2005; Toker et al. 2012], low-ability users might benefit most from adaptive support. Furthermore, Carenini et al. [2014] have shown that the effect of visualization interventions can depend on such characteristics—for example, showing that a user's subjective rating of different highlighting mechanisms is affected by visual working memory. Therefore, adaptive support not only requires identifying users who are currently "struggling" with the task but also consists of predicting and tailoring to each user's individual characteristics.

Second, we are interested in determining which eye gaze features are most informative for classification. As will be shown in this article, different task and user classifications rely on different features, suggesting that adaptive applications need to monitor specific features depending on the intended adaptation purpose. Moreover, the discovery of the most discriminatory features might also provide new suggestions with regard to *how* the system can adapt to support the different tasks and/or user characteristics. For example, if the number of gaze transitions to a certain area of interest (AOI) (e.g., a graph's legend) is found to be very high for users with low cognitive abilities, and knowing that these users are typically less efficient and/or effective on their tasks, we may want to provide help that focuses particularly on reducing the need for visits to this area. Similarly, by finding gaze behaviors that are often exhibited by high-ability users, we might devise adaptive interventions that can encourage low-ability users to also change to such behavior.

This research was first presented in Steichen et al. [2013]. Here, we expand on that work with additional experiments on predicting the task difficulty, as well as the users' expected performance (in terms of completion time). Second, to better evaluate the feasibility of real-time classification, we calculated results at absolute time intervals for each of our experiments, such as the level of accuracy after seeing 5s of gaze data (see Section 5.1). Third, for each of our experiments, we also provide results for a dataset that only includes information on AOIs. Last, in addition to classification accuracy, we also calculated area under receiver operating characteristic (ROC) results—a measure commonly used in machine learning to measure the discriminatory power of models [Egan 1975]—to further strengthen the validity of our findings.

The remainder of this article is structured as follows. First, we provide an overview of related research in adaptive visualization and eye tracking, as well as the most recent findings on the impact of individual user differences in visualization. Next, we present the user study that provided the gaze data for our research. This is followed by a series of classification experiments that we ran on this gaze data to answer the research questions outlined previously. Finally, we conclude with a discussion of the overall findings and outline several directions for future work.

2. RELATED WORK

Adaptation and personalization have long been established as effective techniques to support individual users in a variety of tasks and applications, including personalized search and adaptive hypermedia [Steichen et al. 2012], and desktop assistance and e-learning [Jameson 2008]. By contrast, information visualization research has traditionally maintained a static, one-size-fits-all approach by ignoring an individual user's needs, abilities, and preferences. In particular, early automatic visualization systems have focused only on adapting the visualization to task or data properties that are known a priori [Casner 1991; Mackinlay 1986] rather than dynamically inferring individual properties during visualization usage. An exception to this nonadaptive paradigm is presented in Grawemeyer [2006], where users' visualization expertise and preferences are dynamically inferred through monitoring visualization selections (e.g., how long it takes a user to decide which visualization to choose). Using this inferred level of user expertise and preferences, the system then attempts to recommend the most suitable visualizations for subsequent tasks. Results from the user studies in Grawemeyer [2006] show that the recommendations indeed lead to better user performance in terms of task effectiveness (i.e., accuracy), as well as user efficiency (i.e., time on task). However, this work does not actively monitor a user during a task and thus cannot adapt in real time to help the user with the current task. In contrast, the system developed by Gotz and Wen [2009] actively monitors real-time user behavior during visualization usage to infer needs for intervention. In their work, interaction data (i.e., mouse clicks) are constantly tracked to detect suboptimal usage patterns—that is, activities of users that are of a repetitive (and hence inefficient) nature. Each of these suboptimal patterns indicates that an alternative visualization may be more suitable to the current user activity. The patterns used in their paper include scanning (a user is iteratively inspecting over similar visual objects), flipping (iteratively changing filter constraints), swapping (repeatedly rearranging the order of data dimensions), and drilling (repeatedly filtering down along orthogonal dimensions). Once these patterns are detected, the system triggers adaptation interventions similar to those in Grawemeyer [2006], namely they recommend alternative visualizations that may be more suitable for the current activity (e.g., the location of a set of hotels may be best viewed using a map visualization rather than a user having to repeatedly drill down to this information for each result). However, there are a number of shortcomings of this work. First of all, the usage patterns, as well as the respective visualization recommendations, are determined by experts a priori rather than being based on experimental findings. Second, their system is only able to provide adaptations for visualizations that allow users to interact directly with the visualizations either through mouse clicks or other forms of direct user input. This approach therefore does not work if a user is simply "looking" at a visualization without manipulating its controls/data. Third, their patterns do not try to infer general (low-level) visualization tasks (e.g., filter, compute derived value). Last, their approach does not attempt to adapt to any individual user characteristics.

As mentioned previously, since visual scanning and processing are fundamental components of working with any visualization (they are in fact the only components for noninteractive visualizations), it is important to consider eye tracking as a source of real-time information on user behavior. Although this technology is currently confined to research environments (mostly due to the high cost of eye-tracking devices), the rapid development in affordable, mainstream eye-tracking solutions (e.g., using standard Web cams) will enable the widespread application of these techniques in the near future [Sesma et al. 2012]. In the field of cognitive and perceptual psychology, the use of eye tracking has long been established as a suitable means for analyzing user attention patterns in information processing tasks [Rayner 1998]. Similarly, research in this field has investigated the impact of individual user differences on basic reading and search tasks [Rayner 1995]. More recently, the fields of human-computer interaction and information visualization have also started to use eye-tracking technology to investigate trends and differences in user attention patterns and cognitive/decision processing. This research has typically focused on either identifying pattern differences for different visualizations [Goldberg and Helfman 2011] or task types (e.g., reading vs. mathematical reasoning) [Iqbal and Bailey 2004], or on explaining differences in user accuracy between alternative interfaces [Plumlee and Ware 2006]. However, these studies have generally only attempted to gain insights into differences in gaze behaviors for different tasks and/or interfaces rather than providing a means for directly driving adaptive systems. In particular, these analyses have typically consisted of offline processes that require further human analysis (e.g., manually analyzing eye gaze coordinate plots [Iqbal and Bailey 2004]). In terms of actually using raw eye-tracking data for real-time prediction, most research has so far focused on identifying the user's cognitive processes while she is performing nonvisualization activities, such as during exploratory e-learning [Kardan and Conati 2012; Conati and Merten 2007], quizzes [Courtemanche et al. 2011], simple puzzle games [Eivazi and Bednarik 2011], or information search tasks (e.g., word search) [Simola et al. 2008]. By contrast, our gaze-based work focuses on information visualization, where a user's main activity is to perform simple visualization lookup and comparison tasks.

It is also important to note that none of the preceding approaches has attempted to adapt to user differences other than expertise. However, recent research has shown that other user traits can in fact significantly influence task performance, especially in the field of information visualization. For example, a user's spatial abilities have been shown to influence a user's performance in visual navigation [Chen and Czerwinski 1997] and information search tasks [Westerman and Cribbin 2000].

Similarly, Ziemkiewicz et al. [2011] and Green and Fisher [2010] have looked at the influence of a user's personality traits, showing that locus of control (internal vs. external) can impact visualization performance. Cognitive measures such as perceptual speed and visual working memory have particularly been shown to influence a user's ability to complete basic visualization tasks effectively [Conati and Maclaren 2008; Velez et al. 2005]. For example, it has been shown that users with high perceptual speed have significantly faster completion times and accuracy on certain tasks. These results have been confirmed and extended in a recent study by Toker et al. [2012], where perceptual speed, visual and verbal working memory, and user expertise were shown to influence not only a user's task performance but also satisfaction regarding different visualization types. Most recently, it was found that these individual user differences have an impact on different user eye gaze measures Toker et al. [2013], which directly serves as the motivation for the work in this article on using gaze data to dynamically identify and adapt to user cognitive abilities.

3. USER STUDY

As mentioned in the Introduction, this article is part of our ongoing work on designing user-adaptive information visualizations. In particular, our research studies both the effect that different user characteristics have on visualization performance and the real-time detection of task and user characteristics to be able provide appropriate interventions (the focus of this article). For these purposes, we conducted a user study during which users had to perform a battery of visualization tasks using two alternative basic visualization techniques, namely bar graphs (Figure 1, top) and radar graphs (Figure 1, bottom). By choosing two different types of visualizations, we aimed to investigate the generality of our results.

Bar graphs were chosen because they are one of the most popular and effective visualization techniques. We chose radar graphs because although they are often considered inferior to bar graphs on common information seeking tasks [Few 2005], they are still widely used for multivariate data. Furthermore, there are indications that radar graphs may be just as effective as bar graphs for more complex tasks [Toker et al. 2012].

3.1. Study Tasks

The task domain used in the study required users to evaluate the performance of one or two students in eight different academic courses (using an artificial dataset). We chose this domain to avoid an effect of participant's domain expertise on our results. In the context of this domain, we developed a set of tasks that varied both in task type and task complexity. In terms of different task types, we based our questions on a set of general visualization tasks that had been identified by Amar et al. [2005] to be "representative of the kinds of specific questions that a person may ask when working with a data set." To keep the study conditions manageable, we chose a selection of five task types: *retrieve value* (RV), *filter* (FI), *compute derived value* (CDV), *find extremum* (FE), and *sort* (SO). The types were chosen so that each of our two target visualizations would be suitable to support them. Example questions for each of these task types are shown in Table I.

To vary the task complexity, we differentiated between single and double tasks. Single tasks required participants to compare one student's performance with the class average for the eight academic courses (e.g., "In how many courses is Alice below the class average?"), whereas double tasks required participants to compare the performance of two students with the class average (e.g., "Find the courses in which Andrea is below the class average and Diana is above it."). In total, our study comprised five single tasks, one for each task type (i.e., RV1, FI1, CDV1, FE1, SO1) and four double tasks (RV2,

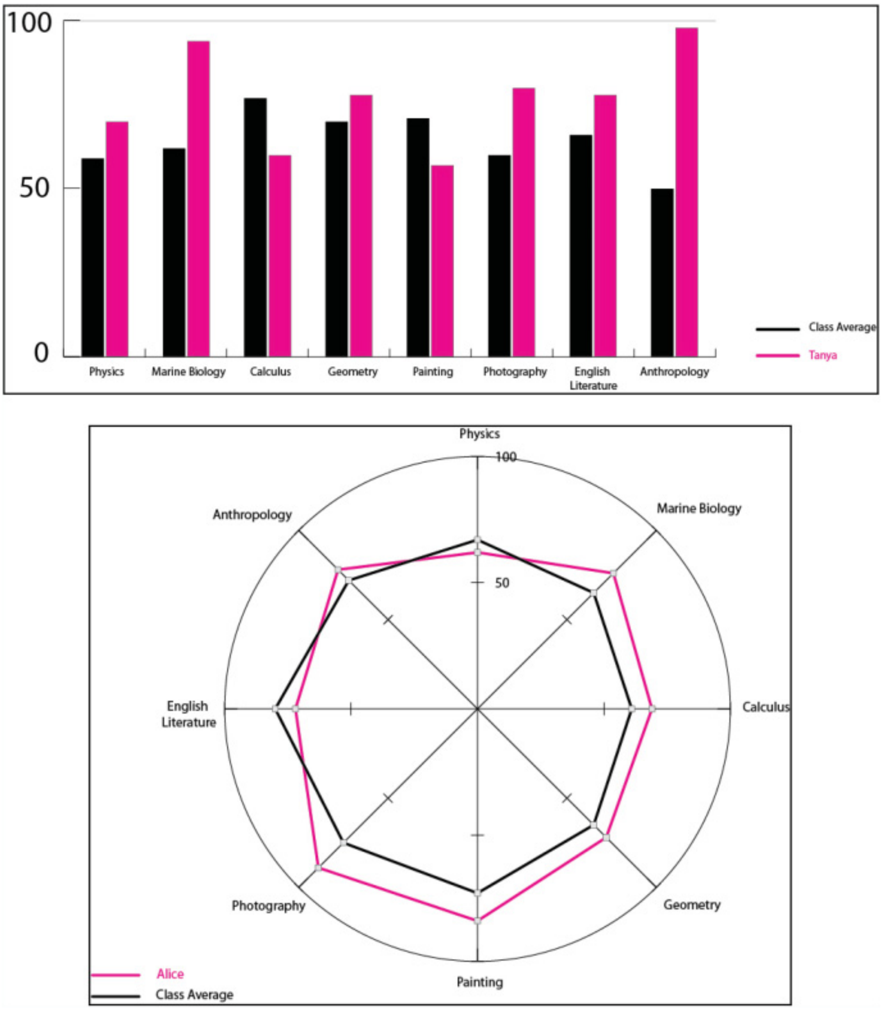


Fig. 1. Sample bar (top) and radar graph (bottom).

Table I. Example Task Questions

RV	Did Alice receive a higher mark in Marine Biology or Painting?
FI	In which course(s) is Mary above the class average? (Select all that apply).
CDV	In how many courses is Alice below the class average?
FE	In which course does Alice deviate most from the class average?
SO	What are Mary's two strongest courses?

CDV2, FI2a, FI2b), meaning that the most fine-grained task type/complexity classification could consist of nine classes (see classification experiments in Section 5.2).

3.2. Cognitive Abilities

The long-term user traits that we investigated in this study consisted of the following three cognitive abilities: perceptual speed (a measure of speed when performing simple perceptual tasks), verbal working memory (a measure of storage and manipulation capacity of verbal information), and visual working memory (a measure of storage and manipulation capacity of visual and spatial information). Perceptual speed and visual working memory were selected because they were among the perceptual abilities explored by Velez et al. [2005], as well as among the set that Conati and Maclaren [2008] found to impact user performance with radar graphs and a multiscale dimension visualizer (MDV). We also chose verbal working memory because we hypothesized that it may affect a user's performance with a visualization in terms of how the user processes its textual components (e.g., legends).

3.3. Study Procedure

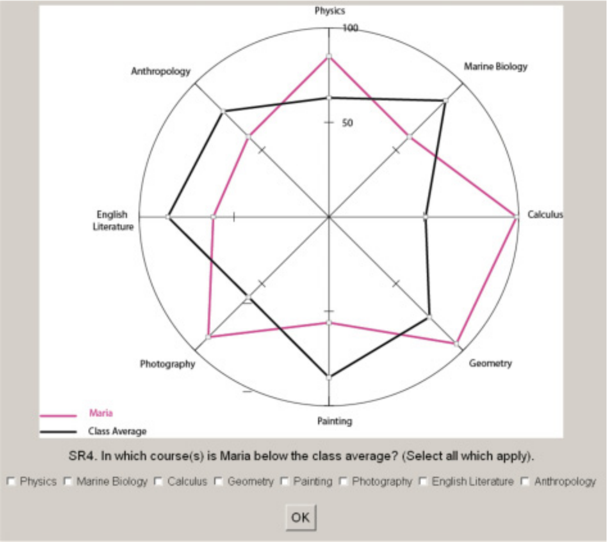
Thirty-five subjects (18 female) participated in the experiment, ranging in age from 19 to 35 years. Participants were recruited via advertising at our university, with the aim of collecting a heterogeneous pool with suitable variability in their cognitive abilities. Ten participants were computer science students, whereas the rest came from a variety of backgrounds, including microbiology, economics, classical archaeology, and film production. The experiment was designed and pilot tested to fit in a single session lasting at most 1 hour. Participants began by completing tests for three cognitive measures: a computer-based OSPAN test for verbal working memory [Turner and Engle 1989] (lasting between 7 and 12 minutes), a computer-based test for visual working memory [Fukuda and Vogel 2009] (10 minutes long), and a paper-based P-3 test for perceptual speed [Ekstrom and U.S. Office of Naval Research 1996] (3 minutes long). The experiment was conducted on a Pentium 4, 3.2GHz, with 2GB of RAM and a Tobii T120 eye tracker as the main display. Tobii T120 is a remote eye tracker embedded in a 17" display, providing unobtrusive eye tracking. After a short calibration of the eye tracker, participants underwent a training phase to familiarize themselves with the two visualizations and study tasks. Participants then performed 14 tasks per visualization—that is, 2 × 5 single and 1 × 4 double (note that each user saw the exact same 28 graphs). The presentation order with respect to visualization type was fully counterbalanced across subjects. Although there was still a remaining training effect (as previously reported in Toker et al. [2012] and recently discussed in Toker et al. [2014]), the counterbalancing ensured that the data was balanced for our classification experiments (albeit containing some potential noise that may have slightly decreased some classification accuracies).

For each task, users were presented with a radar/bar graph displaying the relevant data, along with a textual question (Figures 2 and 3). Participants would then select their answer from a drop-down list, along with their confidence in their answer (between 1 and 5), and click OK to advance to the next task. The experimental software was fully automated and coded in Python.

4. EYE TRACKING MEASURES AND FEATURES

An eye tracker captures gaze information through fixations (i.e., maintaining gaze at one point on the screen) and saccades (i.e., a quick movement of gaze from one fixation point to another), which can be analyzed to derive a viewer's attention patterns. For our experiments, we generated a large set of eye-tracking features by calculating statistics on basic eye-tracking measures (Table II).

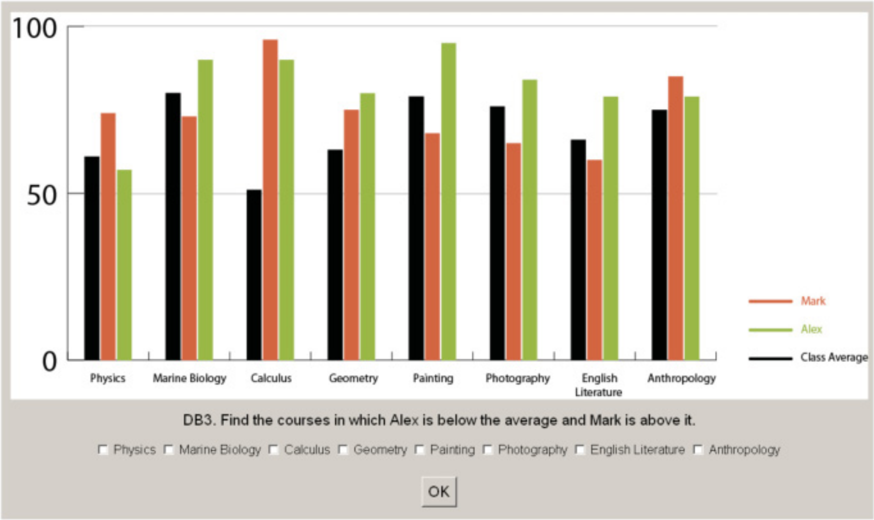
Of these basic measures, fixation rate, number of fixations, and fixation Duration are widely used in eye tracking studies. In addition, we included saccade length (e.g., distance d in Figure 4), relative saccade angle (e.g., angle y in Figure 4) and absolute saccade angle (e.g., angle x in Figure 4), as suggested in Goldberg and Helfman [2010], because these measures are potentially useful for summarizing trends in user attention patterns within a specific interaction window, such as if the user's gaze follows a planned sequence (as opposed to being scattered).



SR4. In which course(s) is Maria below the class average? (Select all which apply).

☐ Physics ☐ Marine Biology ☐ Calculus ☐ Geometry ☐ Painting ☐ Photography ☐ English Literature ☐ Anthropology
OK

Fig. 2. Sample experimental screen of a single-complexity radar graph.



DB3. Find the courses in which Alex is below the average and Mark is above it.

☐ Physics ☐ Marine Biology ☐ Calculus ☐ Geometry ☐ Painting ☐ Photography ☐ English Literature ☐ Anthropology
OK

Fig. 3. Sample experimental screen of a double-complexity bar graph.

Table II. Description of Basic Eye-Tracking Measures

Basic Gaze Measures	Description
Number of fixations	Number of eye fixations detected during an interval of interest
Fixation rate	Number of fixations divided by time interval, e.g., fixations per millisecond
Fixation duration	Time duration of an individual fixation
Saccade length	Distance between the two fixations delimiting the saccade (d in Fig. 4)
Relative saccade angles	The angle between the two consecutive saccades (e.g., angle y in Fig. 4)
Absolute saccade angles	The angle between a saccade and the horizontal (e.g., angle x in Fig. 4)

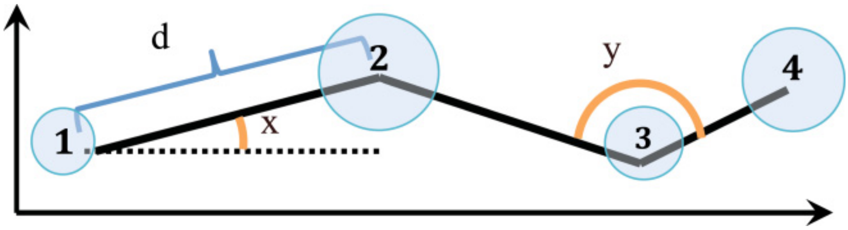
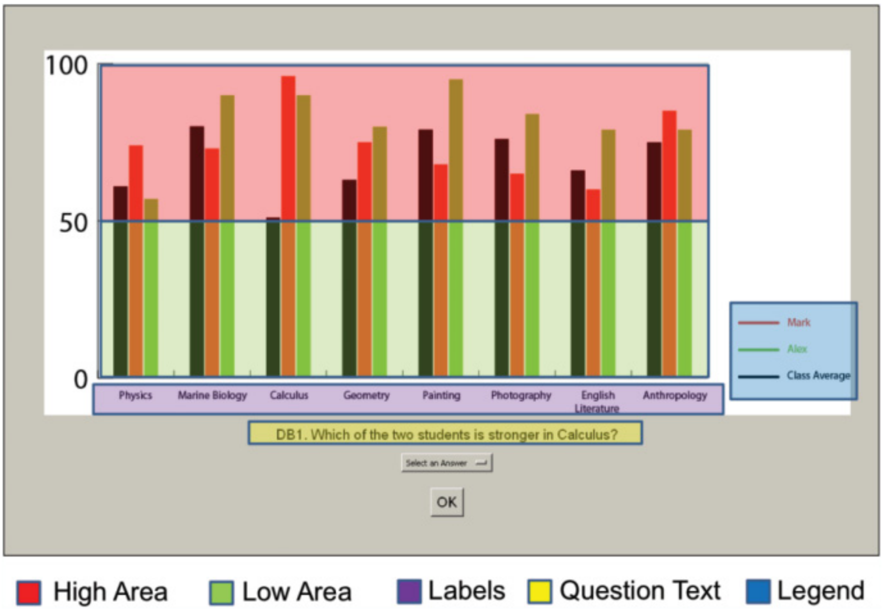


Fig. 4. Saccade-based eye measures.



DB1. Which of the two students is stronger in Calculus?

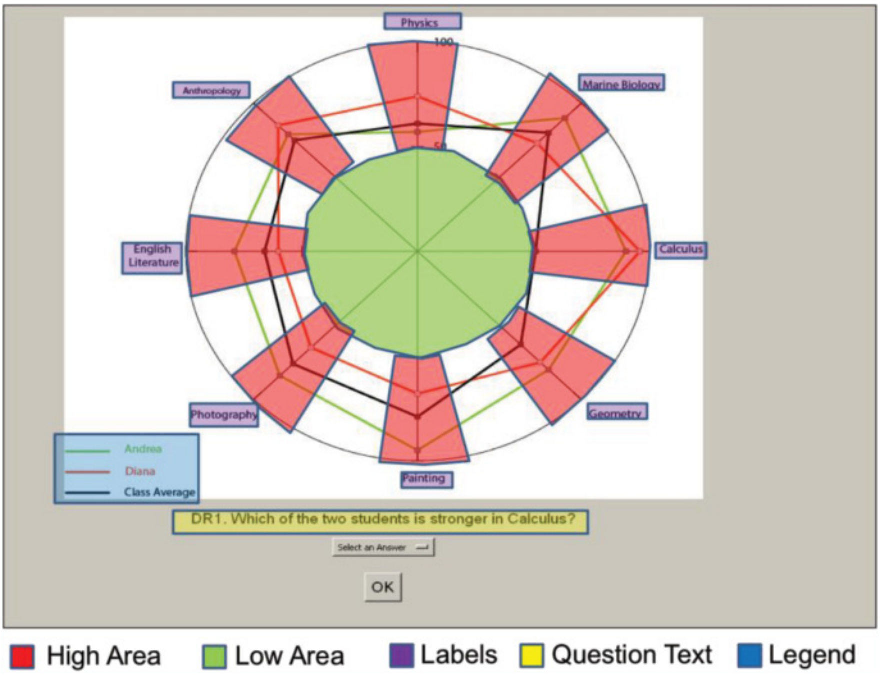
Fig. 5. The five AOI regions defined over a bar graph.

The raw gaze data from the Tobii eye tracker was processed using our open-source data analysis toolkit EMDAT, which is freely available for download and extension by the research community.[1] The toolkit computes statistics such as sum, average, and standard deviation over the eye-tracking measures with respect to (1) the overall screen, to get a sense of the complete interaction with the task (high-level measures from now on) and (2) specific AOIs, identifying parts of the interface relevant for understanding a user's attention processes during each task (AOI-level measures from now on). A total of five AOIs were defined for each of the two visualizations.

These regions were selected to capture the distinctive and typical components of the two visualizations used in the study. Figures 5 and 6 show how these AOIs map onto bar and radar graph components, respectively.

—*High area*: Covers the upper half of the data elements of each visualization. This area is the graphical portion of an Infovis that contains the relevant data values. On the bar graph, it corresponds to a rectangle over the top half of the vertical bars (see Figure 5); for the radar graph, it corresponds to the combined area of the eight trapezoidal regions covering the data points (see Figure 6).

[1] https://www.cs.ubc.ca/group/iui/EMDAT/index.html.



DR1. Which of the two students is stronger in Calculus?

Fig. 6. The five AOI regions defined over a radar graph.

—*Low area*: Covers the lower half of the data elements for each visualization.
—*Labels:* Covers all of the data labels in each graph.
—*Question text*: Covers the text describing the task to be performed.
—*Legend*: Covers the legend showing the mapping between each student and the color of the visualization elements that represent her performance.

The selection of these five AOIs is the result of a trade-off between having detailed information on user attention over areas that are salient for task execution and keeping the number of AOIs manageable for real-time computation. Note that the data values used in the experiment generally ranged between 40 and 100, thereby providing an opportunity to have both "High Area" and a "Low Area" AOIs. If the displayed data values were to range between 0 and 100, there would only be room for a single "Data Value" AOI. However, as will be shown in this article, the Low Area AOI did not play a significant role in our classification experiments; therefore, the results are likely to hold in case there is only one such Data Value AOI.

Overall, a total of 74 features were calculated from the gaze data (Table III). For experimental purposes, we differentiated between a feature set that contained all features, including high-level and AOI features (called the *Full* set from now on), one that did not contain features relating to AOIs—that is, only containing the task-level features (called the *No AOI* set), and one that only contained features relating to AOIs (called the *Only AOI* set). This differentiation was chosen to evaluate the relative information gain attained from AOI and non-AOI features—for instance, how much can be inferred from a user's gaze with and without information on the specific visualization at which the user is looking (similar to what was done in Bondareva et al. [2013] for assessing student learning with an intelligent tutoring system).

Table III. Eye-Tracking Features

HIGH-LEVEL FEATURES
Fixations (2): Number of fixations, fixation rate
Fixation durations (3): Sum, mean, std. deviation
Saccade length (3): Sum, mean, std. deviation
Relative saccade angles (3): Sum, mean, std. deviation
Absolute saccade angles (3): Sum, mean, std. deviation
AOI-LEVEL FEATURES (for each AOI)
Number of fixations in AOI (5)
Sum and mean of fixation durations in AOI (10)
Time to first fixation in AOI (5)
Longest fixation in AOI (5)
Proportion of total number of fixations in AOI (5)
Proportion of total fixation durations in AOI (5)
Proportion of total number of transitions from this AOI to every other AOI (including self-transitions) (25)

5. CLASSIFICATION EXPERIMENTS

The classification experiments described in this section use the previously mentioned features to infer a number of task properties, user performance, and user cognitive traits. In particular, we investigate the extent to which these factors can be inferred from gaze data (research question Q1 in the Introduction), as well as what gaze features are most important for classification (research question Q2).

First, we provide a quick overview of the experimental process used for classification. This is followed by a detailed analysis of each of the classification results, which includes classification accuracy for task type (at different granularities); task complexity; task difficulty; user performance; and accuracy on classifying the three user cognitive abilities of perceptual speed, visual working memory, and verbal working memory. In addition, we ran a classification experiment for predicting the currently active visualization type (i.e., bar graph vs. radar graph) to evaluate the extent to which this information can be inferred when it is not available to the system (i.e., if the visualization system and the eye-tracking component are independent). We conclude with a summary of the overall results, as well as a discussion regarding the extent to which these results could be used for providing adaptive visualizations.

5.1. Experimental Process

Using the gaze features described in the previous section, we generated a number of datasets to simulate partial observation of gaze data during each task. We used two different processes for generating this "over-time" data, each serving a distinct analysis purpose. First, we generated partial observation datasets based on *relative* length, such as the first 10%, 20%, 30%, and so forth, of each trial. The goal of this analysis was to determine if there are observable eye gaze patterns regardless of the actual time it took users to finish the task—for example, if the first 20% of a user's interaction is particularly good at classifying user characteristics. Although this approach can give valuable insights into generalizable trends and patterns, it requires a task to be fully completed to determine what constitutes 100% of the interaction. For this reason, we also generated a second batch of partial observation datasets based on *absolute* length—that is, the first 1,000ms, 2,000s, 3,000ms, and so forth, of each trial. These datasets are more accurate in simulating classification accuracies *while* a user is interacting with (i.e., looking at) the visualization. The goal of this analysis is hence to investigate the feasibility of real-time interventions when integrating the classification component into a live user-adaptive visualization system.

Each of the datasets has a total number of 725 instances, which is a result of pruning the complete set of 980 trials (i.e., 35 subjects × 14 tasks × 2 visitations) to only contain trials with 90% of valid gaze samples.[2]

We used the WEKA data mining toolkit [Hall et al. 2009] for model learning and evaluation. For model learning, we tried a number of different classifier types (Decision Trees, Support Vector Machines (SVMs), Neural Networks, and Logistic Regression) with feature selection and 10-fold cross-validation for model evaluation. We used the standard evaluation metrics of *Accuracy* and *Area under ROC*. In all of our experiments, Logistic Regression (LR from now on) was the classifier with the highest accuracy and ROC. In the following sections, the performance of this classifier is evaluated on the Full, No AOI, and Only AOI datasets. As a baseline for comparison, we use a classifier that always selects the most likely class—for example, for task complexity, the baseline classifier would always predict a task to be single, since there are more single tasks overall (thus failing in all cases of double tasks). Results are generated using the WEKA experiment API with the default 10(repetition) $*$ 10(cross-validation) setting, and statistical significance is tested using t-tests with Bonferroni adjustment on pairwise comparisons between the different classifiers. All reported results are statistically significant (at $p < 0.05$), unless mentioned otherwise. In cases of two-class classifications, we also present the strongest features generated by feature selection. For simplicity, in the case of multiclass classifications, we do not present feature selection in detail, given that this involves presenting n-1 feature selection results (with n = number of classes). Instead, we discuss the impact of features only with respect to the performance of the Full versus No AOI versus Only AOI datasets. In addition, note that in some cases, the y-axis has been readjusted to better show the relative classification performances over time (particularly for the cognitive abilities).

5.2. Classification Results for Task Types and Task Complexity

As explained in Section 3.3, users performed tasks of varying type and complexity. In this section, we first show that task type can be predicted with reasonable accuracy even when tasks are defined at a very fine granularity (in our case comprising nine different task types). Our analysis also reveals that some tasks are frequently confused with one other, suggesting that some fine-grained task types may actually entail similar user strategies. We then present classification results for more coarse-grained task type classifications (five task types, three task types), as well as a two-class classification for task complexity (single vs. double scenario tasks). We show that we can achieve high accuracies for each of these predictions; that a classifier using all eye gaze information generally performs best; and that classification accuracy generally increases after seeing more data.

Task type—nine tasks. The most fine-grained analysis splits tasks into nine different classes, one for each separate question type-complexity combination used in the study, such as *Single Retrieve Value* (RV1) and *Double Retrieve Value* (RV2). Because of the high number of classes, this case represents a difficult multiclass classification challenge, with a baseline classification accuracy of only 15.45% (i.e., always predicting RV1, since this task is, after data pruning, the most common task with 112 out of 725 instances). Nonetheless, when looking at the over-time performance of the LR classifier using the full feature dataset (LR-Full from now on), a classification accuracy of 56.60% can be achieved after seeing all available data (i.e., after 100%) (Figure 7).

[2] Note that gaze data during a trial can be lost due to the subject looking off the screen; due to loss of calibration from rapid movement, blinking, or other such events; or due to blocking of the infrared beam to the user's eyes (e.g., by the user's hands).

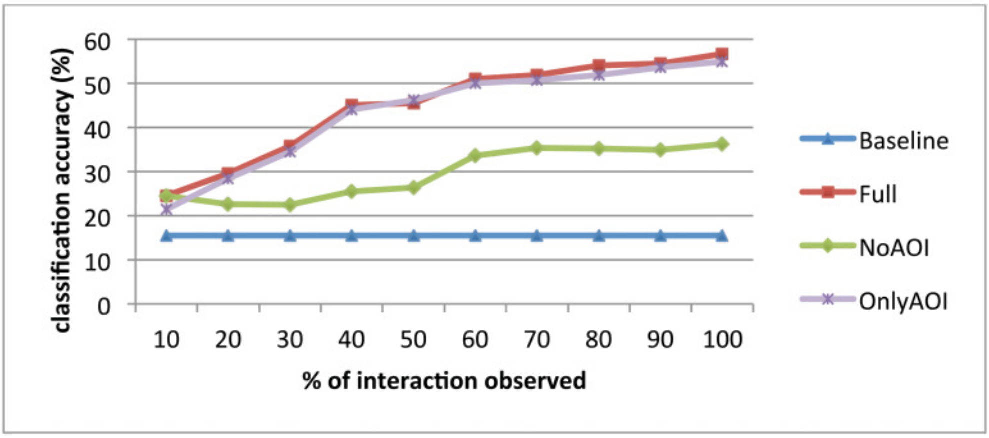


Fig. 7. Task type—classification accuracy (nine tasks).

Table IV. Overall Average Task Classification (Nine Tasks) Accuracy and ROC for Percentage and Absolute Timings

	Baseline	Full	No AOI	Only AOI
Percentage (Accuracy)	15.45	44.81	29.69	43.57
Percentage (ROC)	0.50	0.74	0.68	0.72
Absolute (Accuracy)	15.45	41.65	25.03	40.71
Absolute (ROC)	0.50	0.75	0.67	0.71

As shown in Figure 7, classification accuracy grows continuously as more gaze data becomes available, going higher than 50% after seeing 60% of the data. As shown in Table IV, the average classification accuracy over time for LR-Full is 44.81% and for LR-OnlyAOI is 43.57% (with the difference not being statistically significant). Results are not as good for the LR classifier using the No AOI dataset (LR-NoAOI from now on). The average accuracy over time for this classifier is 29.69%, and its maximum accuracy after seeing all of the data is 30.61%, both statistically significantly lower than the corresponding accuracies for LR-Full. Moreover, the accuracy of the LR-NoAOI classifier is not statistically significantly better than the baseline until after seeing 60% of the data. These differences in performance for the Full versus No AOI versus Only AOI datasets indicate that AOI-related features have a strong impact on classification accuracy for task type at this granularity.

As shown in Figure 8, we found a similar trend when comparing classifiers using the Area under ROC measure, with LR-Full and LR-Only AOI again performing best from 20% of the interaction onward (albeit with slightly different margins compared to the accuracy results). In fact, for almost all of the other experiments described in the following sections, we found similar results for both accuracy and Area under ROC. We will therefore only focus on accuracy results from here onward, except for instances where there was indeed a noticeable difference between the two measures.

In addition to these experiments regarding the percentage of observed interactions, we also test classification performance when using absolute time cut-offs. Both the LR-Full and the LR-OnlyAOI classifiers significantly outperform the baseline classifier from the outset (Figure 9), with accuracies close to 40% after only 5,000ms. However, the LR-NoAOI classifier is only statistically significantly better than baseline after 8,000ms.

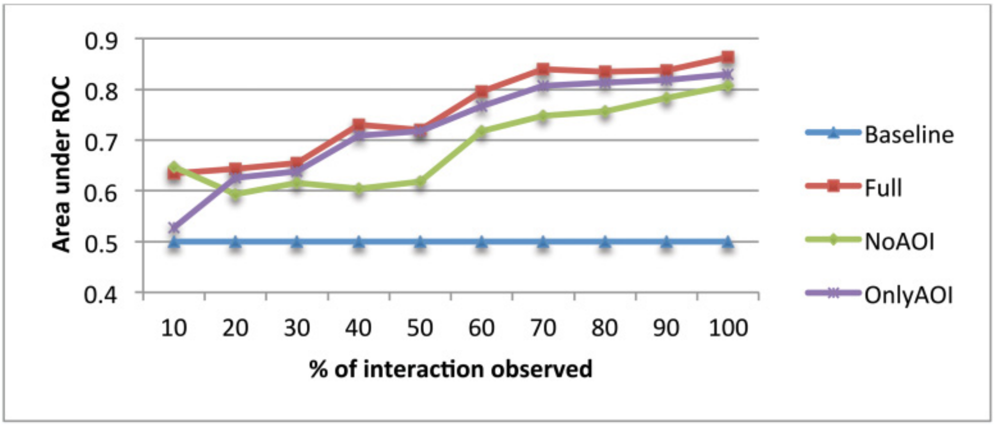


Fig. 8. Task type—Area under ROC results (nine tasks).

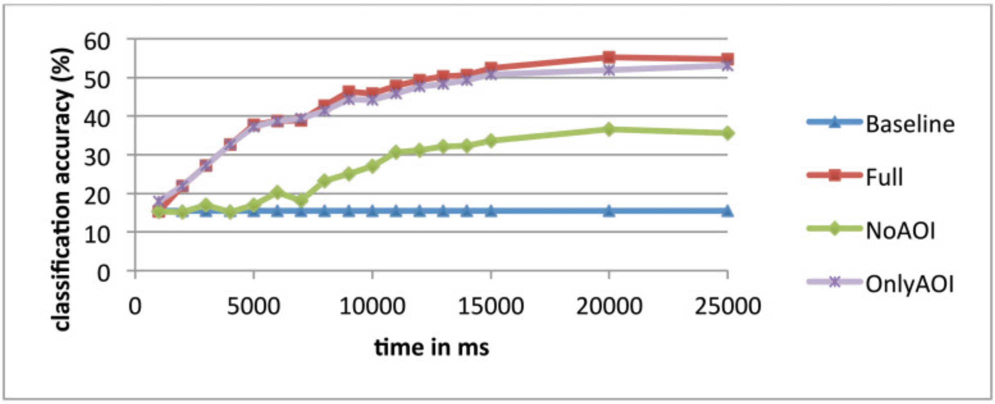


Fig. 9. Task type—classification accuracy at absolute time cut-offs (nine tasks).

When analyzing sources of errors in the confusion matrix, the most commonly confused task pairs were *single filter* (FI1) and *double filter* (FI2) (which is intuitive given the common task type), as well as *single find extremum* (FE1) and *single sort* (SO1) (further discussed in the next section).

Last, as mentioned in Section 5.1, when comparing different types of classifier models (e.g., SVMs, Decision Trees, Neural Networks), we always found LR to yield the highest results (in fact, statistically significantly higher than other models). As an example of this comparison, Figure 10 shows how the various classifiers performed for the nine-task-type classification. As can be seen in this figure, all of the classifiers outperform the baseline from the outset. However, LR performs best across all time intervals. This trend was found across all of our experiments, and we will therefore only discuss the LR results from here onward.

Task type—five tasks. In addition to the fine-grained task analysis involving nine different tasks, we also investigated classifying task type from gaze data when type is defined at a coarser level of granularity that ignores the complexity difference between single and double tasks—for example, ignoring the difference between *retrieve value* when one student is mentioned in the question text (RV1) as opposed to when two students are involved (RV2). Ignoring this difference leaves us with five different

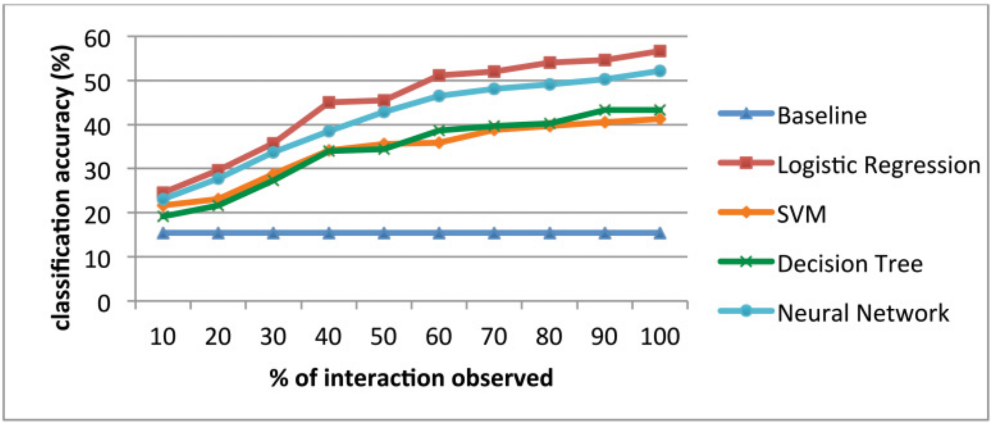


Fig. 10. Comparison of classifier models, each using the Full dataset.

Table V. Overall Average Task Classification (Five Tasks) Accuracy and ROC for Percentage and Absolute Timings

	Baseline	Full	No AOI	Only AOI
Percentage (Accuracy)	27.86	53.36	41.98	51.77
Percentage (ROC)	0.5	0.74	0.67	0.72
Absolute (Accuracy)	27.86	50.00	35.35	48.29
Absolute (ROC)	0.5	0.75	0.66	0.72

classes, corresponding to five different task types from Amar's taxonomy (i.e., retrieve value (RV), filter (FI), compute derived value (CDV), find extremum (FE), and sort (SO)). From the point of view of inferring task type with the goal of providing adaptive interventions specific to tasks types, this five-class classification task is very meaningful, because the classes represent general task types recognized as being common for information visualization.

Similar to the nine-class classification, the trends of the relative and absolute time datasets (i.e., percentage of total observation vs. time in milliseconds) are very comparable. Since the absolute time dataset is more interesting for integration into a live adaptive system (as discussed in Section 5.1.), we only discuss the results of this analysis from here onward. LR-Full reaches an average accuracy of 53.36% over time (Table V), an accuracy of 50% after 8,000ms, and a maximum accuracy of 63.32% after seeing all data. LR-Full statistically significantly outperforms the baseline's accuracy (27.86%, i.e., always choosing filter, the most common task with 202 instances) from the start. As was the case with nine tasks, LR-OnlyAOI is not statistically significantly different from the LR-Full classifier. Similarly, removing AOI-related features statistically significantly reduces accuracy, as shown by the performance of LR-NoAOI in Figure 11. Moreover, this classifier only starts to be statistically significantly better than baseline after 3,000ms.

Although the accuracies are clearly improved over the nine-task classification (as to be expected due to the reduction of the classification complexity), it is still arguable if they are acceptable for real-time fully adaptive interventions. However, one could certainly consider using such a classifier in a mixed-initiative system, where a range of task-appropriate interventions could be recommended to a user rather than applied automatically. This would still potentially reduce a user's workload while not interrupting a task with inappropriate interventions. In addition, it is worth noting that

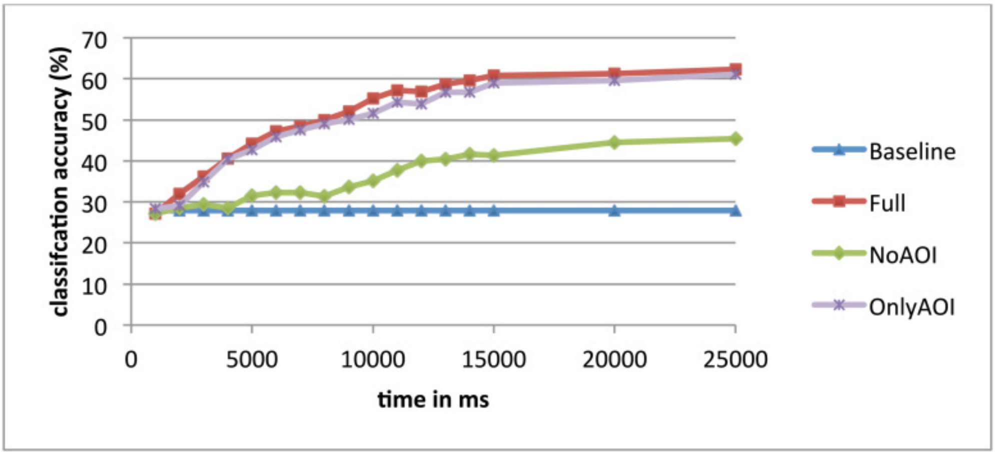


Fig. 11. Task type—classification accuracy (five tasks).

the predictions are solely based on eye gaze data, and that the integration of other data sources (e.g., interaction data) could probably complement and improve results.

When analyzing sources of errors in the confusion matrix, we found two pairs of tasks that are most often confused with one other. The first pair involves the tasks compute derived value (CDV) and filter (FI). For example, in 57% of the cases where CDV was misclassified, the predicted class was FI. This result is not surprising, since both of these tasks essentially involve applying a filter to all data values (e.g., finding values above a given threshold), with the difference being that CDV requires an additional computation (e.g., "In how many courses is student X above the class average?"). Thus, FI can be regarded as a subtask of CDV for the questions used in our study. In fact, as noted by Amar et al. [2005], the filter task "is used as a subtask in many other questions." Adaptations that particularly support this FI task may therefore also be of use to CDV tasks if they contain such a subcomponent. The second pair of tasks often confused with each other involves find extremum (FE) and sort (SO). For example, in 38% of the cases where FE was misclassified, the predicted class was SO. This result is again not surprising given the nature of these two tasks. FE involves going through all values to find the highest value(s) from a set of values, whereas SO involves sorting all values from highest to lowest. Thus, FE essentially involves a subpart of the steps necessary to perform an SO task. This finding confirms the observation in the taxonomy of Amar et al. [2005] that "sorting is generally a substrate for extreme value finding."

The aforementioned relations between the two pairs of frequently confused tasks suggest that combining each pair into one new task type and building a classifier that can recognize this combined type is still valuable for adaptation, since adaptations could be provided to support the common subtask. Thus, in the next section, we evaluate the accuracy of a classifier for task type that involves three classes: FI-CDV (combined), SO-FE (combined), and RV.

Task type—three tasks. When considering only three different task types, LR-Full reaches an average accuracy of 68.42% over time and an accuracy of 70% after only 8,000ms. LR-Full statistically significantly outperforms the baseline's accuracy (48.14%) from the start.

As was the case with nine and five tasks, the performance of LR-OnlyAOI is very similar to the performance of LR-Full. Similarly, removing AOI-related features once again statistically significantly reduces accuracy, as shown by the performance of

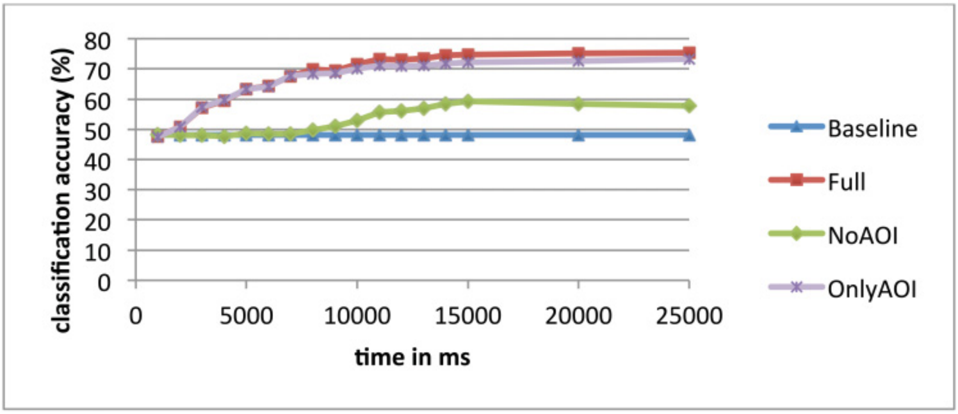


Fig. 12. Task type—classification accuracy (three tasks).

Table VI. Overall Average Task Classification (Three Tasks) Accuracy and ROC for Percentage and Absolute Timings

Percentage (Accuracy)	48.14	68.42	56.75	67.00
Percentage (ROC)	0.5	0.81	0.65	0.80
Absolute (Accuracy)	48.13	67.10	52.56	65.91
Absolute (ROC)	0.5	0.80	0.58	0.80

LR-NoAOI in Figure 12 and Table VI. This classifier only reaches an average accuracy of 56.75% over time and only statistically significantly outperforms the baseline after 10,000ms.

With accuracies reaching 70% after only 8s, one can certainly envision the use of such a classifier to assist a user in an adaptive visualization system. Interventions could consist of recommendations, as well as direct automatic visualization adaptations. However, to validate the practicality of this approach, we will need to run further user studies with a fully implemented adaptive system.

Task types—summary of results. In summary, we found that across all task type granularities, LR with the Full dataset outperformed both the baseline and LR with the No AOI dataset, but not the LR-OnlyAOI classifier, showing the importance of having AOI-related features for task-type classification. Figure 13 summarizes the results in terms of average accuracy over time. As expected, accuracy for all of the classifiers increases as task granularity gets coarser. Although only the classification of three tasks with the LR-Full classifier reaches accuracies that may be suitable for providing reliable task-based interventions, we see these results as being very important for two reasons. First, as we argued earlier, suitable interventions can be provided even if task type is recognized at this coarser level. Second, our results have been obtained by using relatively simple eye gaze features that do not capture gaze patterns beyond simple transitions between two AOIs. Using more complex gaze patterns or additional sources of information to guide classification (see discussion in Section 5.6), it is likely that we can increase accuracy on all of our classification tasks.

Task complexity. The classifier in this experiment predicts if the user is attending to a task of the single or double scenario. As discussed in the Section 3, this distinction provides a measure for task complexity. LR-Full and LR-OnlyAOI are still the most accurate classifiers, with statistically significantly higher average accuracy (80.39%/80.22%) over time than both LR-NoAOI (74.76%) and the baseline classifier

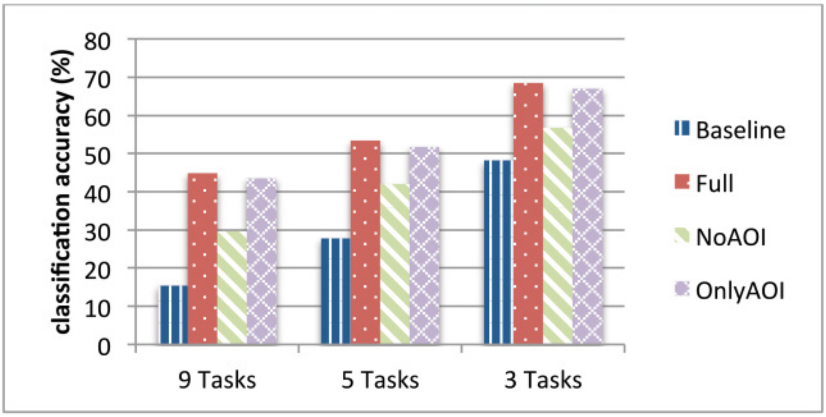


Fig. 13. Task type—average classification accuracy over time for different task granularities.

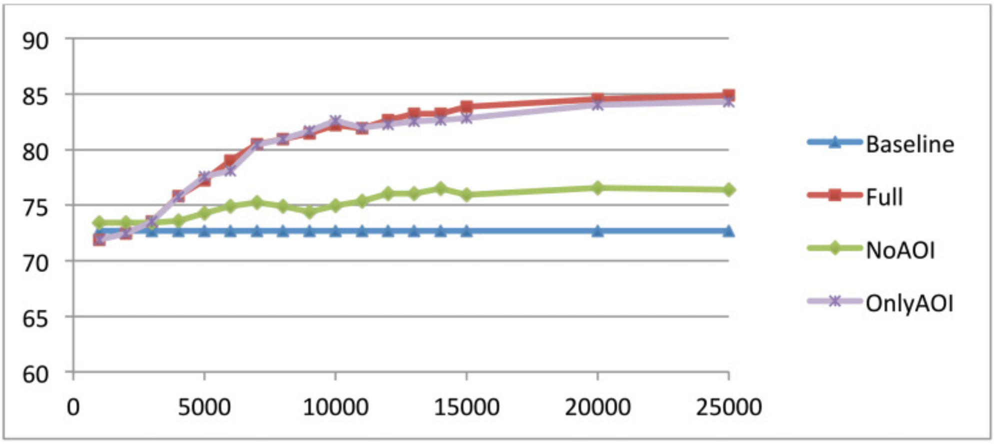


Fig. 14. Task complexity—classification accuracy.

Table VII. Overall Average Complexity Classification Accuracy and ROC for Percentage and Absolute Timings

	Baseline	Full	No AOI	Only AOI
Percentage (Accuracy)	72.69	80.39	74.76	80.22
Percentage (ROC)	0.50	0.81	0.73	0.80
Absolute (Accuracy)	72.69	79.96	75.02	79.74
Absolute (ROC)	0.50	0.81	0.69	0.80

(72.69%, i.e., always choosing single, since this is the dominant class with 527 out of the 725 instances). The top classifiers, LR-Full and LR-OnlyAOI, are once again not statistically significantly different from each other. It should be noted that at 72.69%, the baseline accuracy is relatively high in this experiment, since users performed more than twice as many single tasks than double ones. Nevertheless, all three LR classifiers performed higher, with accuracies reaching up to 84.45% for the Full dataset (Figure 14 and Table VII). Accuracy again improved with more data being observed, and each of the feature sets outperformed the baseline after relatively low amounts of observed data.

Since task complexity consists of a simple two-class classification, we also investigated which specific features from the feature selection process are contributing the most to the classification (note that for multiclass logistic regression involving n classes, this analysis would have been too cumbersome because it would need to consider n-1 feature selection results). The most predictive features were "proportion of total fixation durations in legend AOI," "sum of fixation durations in legend AOI," and "proportionate number of transitions from/to legend AOI to/from high AOI." With increased task complexity, we found that the use of the graph legend increased considerably, both in terms of proportion of total fixation durations (compared to all other AOIs), as well as in terms of transitions (i.e., there were more transitions to and from the legend). This result shows that an increase in data series has an effect on how much users may need to refer back to the legend during a visualization task, as to be expected. Nevertheless, it is an interesting finding that such an increase in complexity can be captured in real time using simple eye gaze measures, which may in turn allow a user-adaptive system to provide adaptations for more complex tasks (e.g., provide support for better legend access and processing).

5.3. Classification Results for Task Difficulty

In addition to "task complexity," which we defined based on the number of data series, we also tried to predict the overall "difficulty" of a task, which we defined based on a combination of subjective and objective measures. For this measure, we again found similar relative performances of the different classifiers, and that accuracy generally improves with more data. We will now first describe in detail how we generated a difficulty value for each task, followed by the detailed results of the classification experiments.

Definition of task difficulty. Defining tasks as being easy or difficult a priori is challenging, since difficulty depends on user expertise and perceptual abilities, which were varied on purpose in our study. We therefore defined task difficulty a posteriori, based on four different measures (two objective and two subjective) aggregated using a principal component analysis (PCA). Because there was a ceiling effect on task correctness, our first objective measure of task difficulty is task completion time (assuming that, in general, more time is needed for more difficult tasks). However, longer completion times may also simply be an indication of a task being longer while not necessarily being more difficult. Therefore, our second objective measure of difficulty is the standard deviation of completion time for each task across all users. A high value of this metric indicates a high variability among users' completion times, an indicator that the task may be difficult or confusing for some users.

Our two chosen subjective measures of task difficulty are based on the users' reported confidence of their performance, which was elicited after each task. The first subjective measure is the average confidence reported by users on each task. Intuitively, less difficult tasks would have higher values for this average. However, we also want to take into account that some users may tend to be more confident overall than other users. Therefore, our second subjective measure is the average deviation of confidence for each task across all users and is computed as follows. For each user, we look at their average confidence across their tasks. Then, for each task, we compute the deviation of confidence as the difference between the user's reported confidence for that task and the user's average confidence across tasks. Finally, for each task, we average the deviation of confidence across all users. This average indicates for which tasks users were giving confidence ratings that were above or below their typical rating.

To combine the preceding four variables, we performed a PCA, which is a form of dimensionality reduction that allows one to identify and combine groups of interrelated

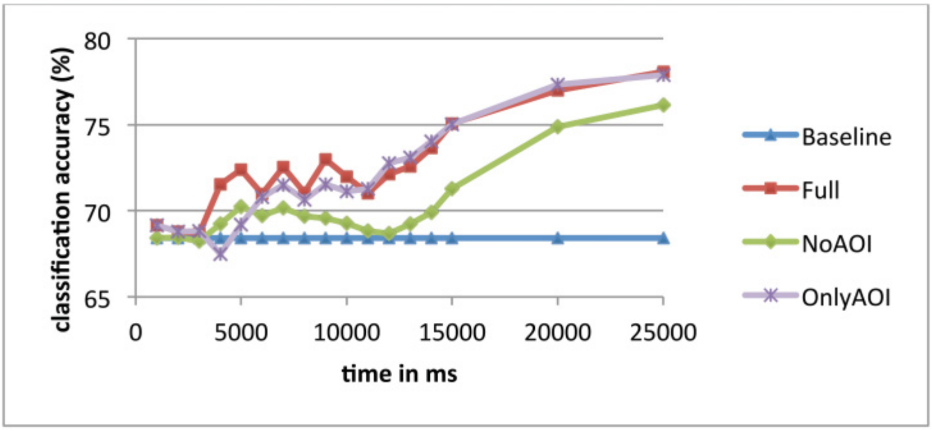


Fig. 15. Task difficulty—classification accuracy.

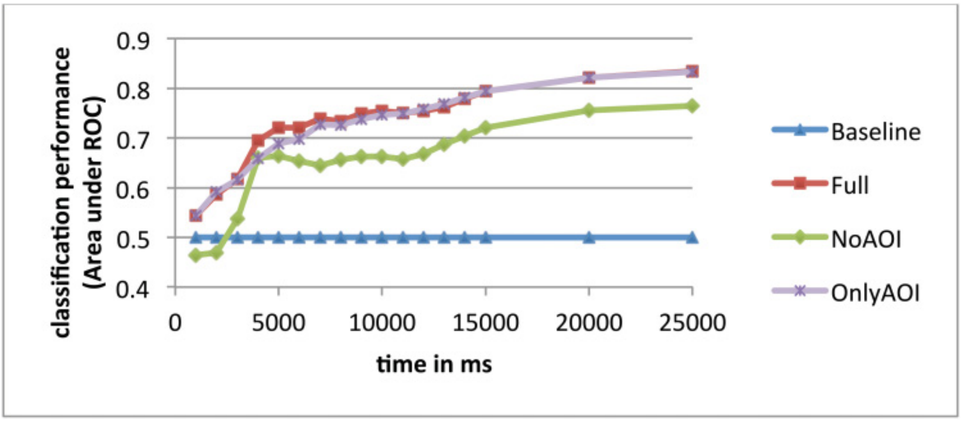


Fig. 16. Task difficulty—classification results using the Area under ROC measure.

variables into components more suitable for data analysis. A PCA on our four measures of task difficulty resulted in one output component. Bartlett's test of sphericity ($x^2 = 73.35$, df $= 6$, $p < .001$) indicated that the PCA was appropriate. Kaiser's sampling adequacy was 0.55, and all variables showed a communality >0.52, which was above the acceptable limit of 0.5. The component that we generated had an eigenvalue over Kaiser's criterion of 1 and explained 62.22% of the variance. In summary, we used the output component generated by this PCA (i.e., dimensional reduction) as the measure of task difficulty, and for classification purposes, we labeled tasks with a negative component as *easy* and tasks with a positive component as *difficult* (resulting in 497 easy and 228 difficult trials). Each of the easy/difficult classes included both bar and radar graph trials, as well as both single and double task trials, thereby showing that difficulty was not solely confined to radar graphs and/or double complexity tasks.

Classification results. Our classification experiments on this task difficulty property showed slightly different patterns (Figures 15 and 16) compared to task complexity—that is, a nonsmooth curve between 5,000ms and 10,000ms. However, as noted in Provost et al. [1998], for datasets that are imbalanced (as was the case for task difficulty), accuracy is less reliable. When analyzing the Area under ROC results, we

Table VIII. Overall Average Difficulty Classification Accuracy and ROC for Percentage and Absolute Timings

	Baseline	Full	No AOI	Only AOI
Percentage (Accuracy)	68.41	76.02	74.46	75.78
Percentage (ROC)	0.50	0.79	0.75	0.78
Absolute (Accuracy)	68.41	72.12	70.36	71.56
Absolute (ROC)	0.50	0.73	0.65	0.72

found that the curve for this measure was indeed relatively smooth (see Figure 16) and that the LR-Full and LR-OnlyAOI datasets once again statistically significantly outperformed LR-NoAOI, as well as the baseline (Table VIII). Relatively high accuracies/Area under ROC are reached after only 5,000ms (72%/0.68) and reach up to 78%/0.86 for LR-Full and LR-OnlyAOI. This classifier could hence be used in an adaptive system to support users during “hard” tasks. However, there is still significant room for improvement in terms of the accuracies, especially compared to the baseline (which could potentially come from different data sources, such as input devices).

In terms of “how” to provide this support, it is again worth investigating the feature selection results to see which particular eye gaze behaviors are most indicative of a user facing a difficult task. The three most predictive features for task difficulty were “proportionate number of transitions from labels to high AOI,” “sum of fixation duration in label AOI,” and “proportion of total fixation durations in high AOI.” These findings indicate that difficult tasks might require users to do more repeated visits of the actual graph values (and their associated labels), perhaps due to a task being ambiguous or simply requiring more cognitive effort. Therefore, if in addition to determining the task type our classifier flags the current task as difficult (e.g., due to the user performing repeated visits to the high AOI area, and/or spending a high proportionate amount of time in this area), the system should try to provide assistive support to the user through an adaptive intervention that reduces this effort.

5.4. Classification Results for User Performance

In addition to predicting what particular task a user is performing (as well as the task’s complexity/difficulty), we also aimed at predicting “how well” a user is/will be performing on this current task. The reasoning for this experiment is that while there might be situations where we could provide adaptations purely based on a predicted task type and/or task characteristics, an adaptive visualization system is arguably most appropriate when a user is currently performing “suboptimally.” In particular, since adaptive interventions could introduce slight disruptions in a user’s workflow, it might be best to only provide interventions when we detect a “slow” user (while ignoring users who are already performing well). To estimate a user’s current performance, we hence are trying to predict if a user’s completion time will be above or below the median time (based on all users) for this particular task. Overall, our results showed that performance can be predicted well within the first stages of the prediction and that more general features (i.e., not specific to any AOIs) are particularly discriminative.

Figure 17 shows the overall classification results based on the 725 valid trials, of which 357 were labeled as “slow” and 368 were labeled as “fast” (note that we did not run separate classification experiments per task). As can be seen in this figure, LR-Full achieves the highest overall accuracies, reaching 65% after only 5,000ms. Overall, all three classifiers are closely matched, with an average accuracy of 65% for LR-Full and 62% for both LR-NoAOI and LR-OnlyAOI (baseline 51%) (Table IX). Nonetheless, the differences observed in the early stages (at 5,000ms) are statistically significant, as are all performances compared to the baseline (from 2,000ms onward).

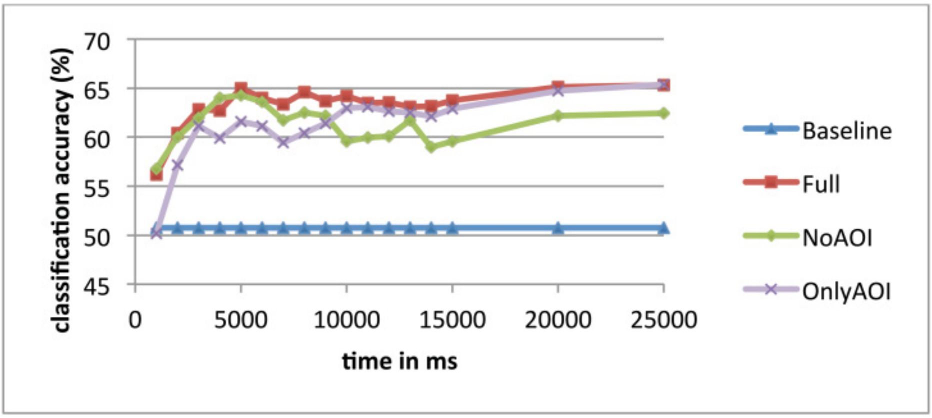


Fig. 17. User performance—classification accuracy.

Table IX. Overall Average Performance Classification Accuracy and ROC for Percentage and Absolute Timings

	Baseline	Full	No AOI	Only AOI
Percentage (Accuracy)	50.75	65.20	62.45	62.23
Percentage (ROC)	0.50	0.73	0.69	0.69
Absolute (Accuracy)	50.75	63.20	61.27	61.09
Absolute (ROC)	0.50	0.68	0.66	0.66

Interestingly, LR-NoAOI outperforms LR-OnlyAOI in the early stages of a user's task, showing that a user's overall performance may be predicted using features that are independent of any AOIs. When analyzing feature selection results for LR-Full at this early stage, we found that "mean saccade length," "standard deviation of absolute saccade angles," and "standard deviation of saccade length" were most predictive (each being non-AOI features). When implementing an adaptive intervention system that targets the early phases of a user's interaction, it may therefore be sufficient to base the classification on non-AOI related features.

However, after 10,000ms, it is once again the LR-OnlyAOI classifier that closely matches the LR-Full classifier, showing that AOI-related information positively contributes to classification accuracy in the later stages. The most predictive features in these stages were "proportion of total number of fixations in text AOI," "proportion of total fixation durations in text AOI," and "proportionate number of transitions from text to low AOI," indicating that users with lower performances (in terms of time) refer more often to the textual information associated with the graph (which could consist of the graph's caption).

5.5. Classification Results for Cognitive Abilities

In this section, we discuss classification results relating to inferring a user's level of visual working memory, verbal working memory, and perceptual speed. The specific task of each of the three classifiers is to infer if a user belongs to either the High or Low category for that measure (based on a median split).

In general, we found similar results across each of these three classification experiments. First, we found that AOI features are again very useful for predictions and that most classifiers outperform the baseline from the outset. Although average accuracies for even the best classifier were rather low (between 56% and 60%; Tables X, XI, and XII), it has to be noted again that these experiments are solely based on simple

Table X. Overall Average Visual Working Memory Classification Accuracy and ROC for Percentage and Absolute Timings

	Baseline	Full	No AOI	Only AOI
Percentage (Accuracy)	54.90	57.47	56.78	56.95
Percentage (ROC)	0.50	0.60	0.57	0.59
Absolute (Accuracy)	54.73	56.93	54.89	57.21
Absolute (ROC)	0.50	0.59	0.52	0.59

Table XI. Overall Average Verbal Working Memory Classification Accuracy and ROC for Percentage and Absolute Timings

	Baseline	Full	No AOI	Only AOI
Percentage (Accuracy)	52.14	60.75	55.40	59.83
Percentage (ROC)	0.50	0.65	0.58	0.64
Absolute (Accuracy)	52.14	60.33	55.84	59.82
Absolute (ROC)	0.50	0.65	0.57	0.64

Table XII. Overall Average Perceptual Speed Classification Accuracy and ROC for Percentage and Absolute Timings

	Baseline	Full	No AOI	Only AOI
Percentage (Accuracy)	50.07	57.07	53.19	57.12
Percentage (ROC)	0.50	0.59	0.54	0.59
Absolute (Accuracy)	50.07	56.29	52.45	56.35
Absolute (ROC)	0.50	0.58	0.53	0.58

eye-tracking measures, which may be improved using additional sources of information (see overall result discussion in Section 5.6). Second, we made several interesting observations when analyzing the accuracies at different data cut-off points. In particular, for each of the experiments, the peak accuracies were actually found during the early stages of each trial, as opposed to after all data had been observed (as found in most of the other classification experiments described earlier). This pattern suggests that a user's cognitive abilities most strongly affect a user's gaze patterns during the initial phase of a visualization task (as shown in Figures 18, 19, and 20) and that these patterns are increasingly "diluted" by other factors (e.g., task type) as the task goes on. Although this goes against the intuition that more data generally helps classification, the analysis of feature selection actually provided some sensible explanations for this finding (discussed next).

For *visual working memory*, the peak accuracy of 60% occurred after 6,000ms (see Figure 18). When analyzing the features that received the highest coefficient during feature selection, we found that the time to first fixation for *text, label*, and *high AOIs* played an important role in classifying users. We found that high visual working memory users had lower times to first fixation (indicated by a negative coefficient), meaning that they were very quick at scanning the various AOIs of the visualization.

Similarly, for *verbal working memory*, the highest classification accuracy for both LR-Full (64%) and LR-OnlyAOI (61%) was found after observing only 3,000ms (see Figure 19).

When analyzing the feature selection results for LR-Full, we found that features related to the *text* and *label AOI* most strongly contributed to the classification accuracy. In particular, high verbal working memory users spent less time in the text AOI, both overall and in proportion to other AOIs. Since users are most likely to read the question text at the beginning of each task, it therefore seems intuitive that the highest accuracies were found after only a few seconds of the data had been observed.

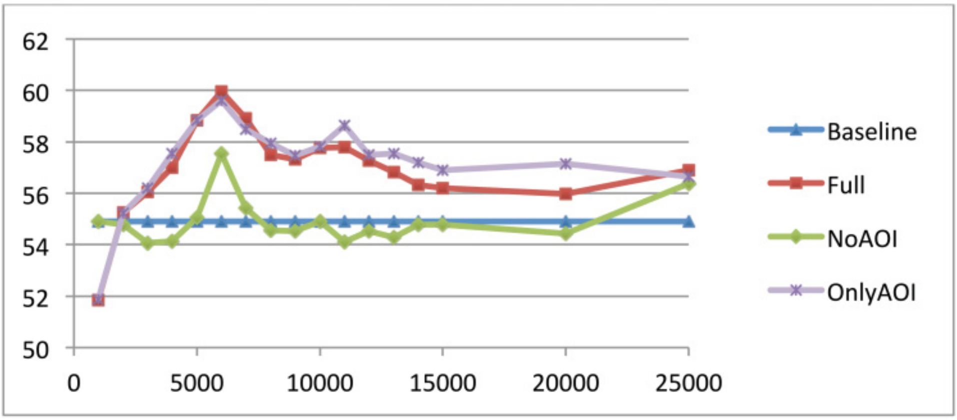


Fig. 18. Visual working memory—classification accuracy for relative time slices.

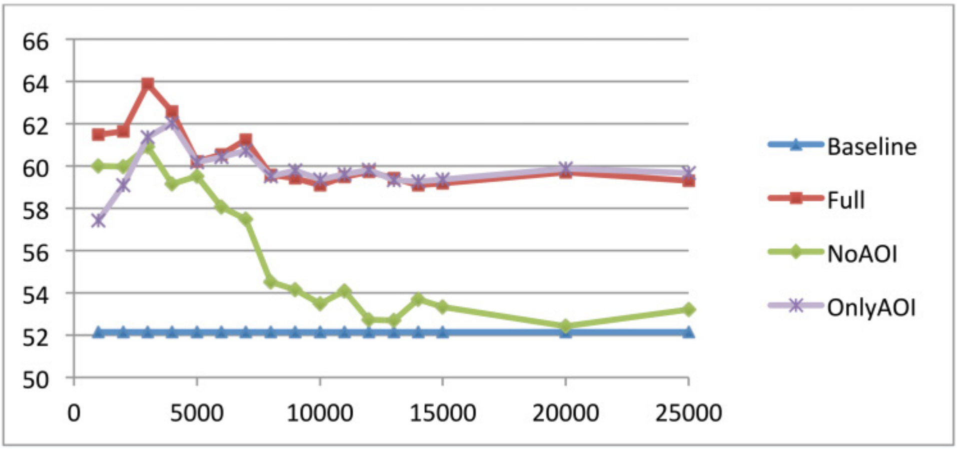


Fig. 19. Verbal working memory—classification accuracy.

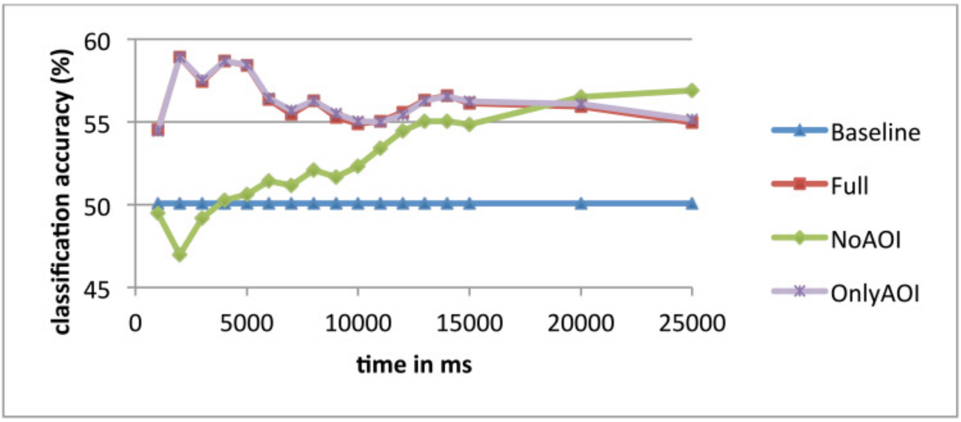


Fig. 20. Perceptual speed—classification accuracy for absolute time slices.

Table XIII. Overall Average Visualization Type Classification Accuracy and ROC for Percentage and Absolute Timings

	Baseline	Full	No AOI	Only AOI
Percentage (Accuracy)	52.69	83.45	66.58	80.39
Percentage (ROC)	0.50	0.91	0.73	0.88
Absolute (Accuracy)	52.69	84.31	64.63	81.67
Absolute (ROC)	0.50	0.91	0.69	0.88

A similar pattern was observed for the *perceptual speed* classification experiments, where the highest accuracy for LR-Full (59%) was found after only 2,000ms of data had been observed (see Figure 20). Interestingly, the LR-NoAOI classifier outperformed LR-Full and LR-OnlyAOI after 20,000ms.

When analyzing the feature selection results during the early stages, we found that features related to the *label* and *legend AOIs* had the strongest coefficients. In particular, we found that high perceptual speed users had a “lower number of fixations in the legend AOI” and “lower proportion of total number of fixations in the legend AOI.” This finding may indicate that low perceptual speed users would benefit from adaptations relating to this particular AOI (through highlighting, facilitating easier access, etc.). In addition, low perceptual speed users had a longer “longest fixation in the label AOI” and a lower “fixation rate” compared to high perceptual speed users. Again, this may indicate that we can provide adaptations particularly tailored toward the label AOI—for example, by temporarily increasing the size of relevant labels to support low perceptual speed users.

However, during the later stages, we found that in addition to fixation rate, the No AOI features of “mean fixation duration” and “sum of absolute saccade angles” were the most predictive (hence explaining the better performance of the LR-NoAOI classifier).

Visualization type. As shown in almost all of the preceding classification results, the inclusion of AOI-related features is critical toward generating predictions for task properties, user performance, and user cognitive measures. However, having these AOI-related features requires knowing which visualization is currently active. Although there are many scenarios in which this information is indeed available to an adaptive component (i.e., if the adaptation component is part of the visualization system), this is not always the case. For example, if an adaptive component were to run as a standalone system in parallel to a separate visualization system (in the context of an information retrieval task when the user gets back a visualization, or in case the adaptive component acts as a complement to a third-party analysis tools, etc.), it would first be necessary to infer the currently active visualization type to utilize the right AOIs for accurate task/user classifications. Thus, in this section, we present results on whether visualization type can be inferred from gaze data. Since AOI information would not be available for this task, LR-Full and LR-OnlyAOI are not applicable in a realistic scenario. For the LR-NoAOI classifier, the average accuracy is 66.58%, which is statistically significantly higher than the baseline (52.69%) (Table XIII). As shown in Figure 21 (note that LR-Full and LR-OnlyAOI are included for completeness), the accuracy of LR-NoAOI continuously grows as more gaze data is observed, reaching 66% after 5,000ms and leveling off at around 70% after 2,0000ms. All accuracies are statistically significantly higher than baseline after only 3,000ms. Although these results are encouraging, further research needs to be conducted in terms of improving accuracies to employ these techniques in a live system (see discussion in Section 5.6).

Regarding the feature selection for this NoAOI classifier, we found that users have different viewing patterns in terms of path angles. Specifically (and not surprisingly), users have more horizontal viewing patterns in the bar graph (lower mean absolute

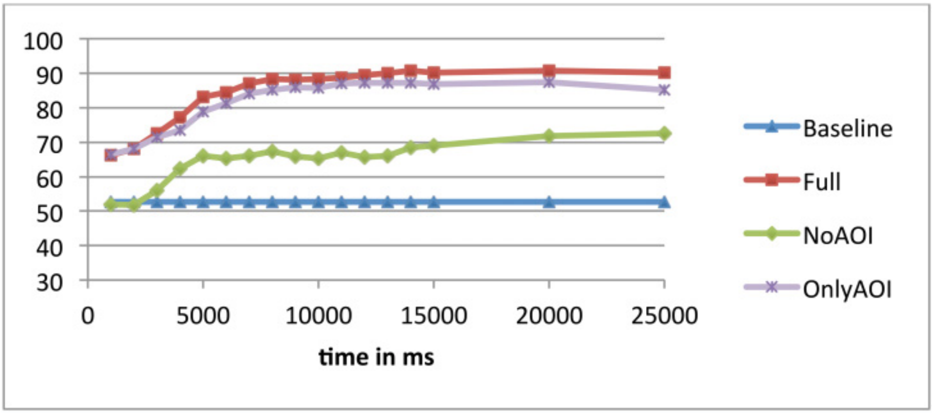


Fig. 21. Visualization type—classification accuracy.

path angles) and more "erratic" saccades (higher standard deviation of absolute path angles), whereas in the radar graph users follow a circular trajectory to view the various data points (indicated by a higher mean of absolute path angles) and have more uniform saccades due to the proximity of the labels to the respective data points (indicated by a lower standard deviation of absolute path angles).

5.6. Summary of Results and Discussion

As outlined in the Introduction, the specific goals of our experiments were to investigate the extent to which a user's current visualization task properties, a user's performance, and a user's long-term cognitive abilities could be inferred solely based on eye gaze data (Q1), as well as which gaze features would be the most informative (Q2). By running a number of classification experiments and analyzing in detail the effects of different feature sets, we have found several interesting results regarding these research questions.

We found that a user's eye gaze behavior provides evidence about each of the visualization task types and characteristics, user performance, and user cognitive abilities. In particular, we showed that for each classification task, gaze-behavior–based predictions outperform a baseline classifier (except for user expertise, which we hence did not discuss further) (Q1). Moreover, we show that for most of the predictions, the classification accuracy is statistically significantly higher even after only partial data observations. We have shown that for some experiments, accuracy is actually highest at the beginning of each task, indicating that a user's eye gaze at this time may contain the most relevant information regarding the target characteristics. These results provide very encouraging evidence that user eye gaze behavior could indeed be used for driving adaptive systems, particularly given that the experiments used a relatively simple set of features. Interestingly, LR consistently achieved the highest accuracies compared to other machine learning models, such as SVM or Decision Trees. Although we do not have an intuitive explanation for this finding, several other works have similarly found LR to perform well with eye gaze data [Kardan and Conati 2012; Bondareva et al. 2013].

It may be argued that from a practical point of view, some of the accuracies are not sufficiently high to be exploited in a live system. In particular, the accuracies relating to the cognitive abilities yielded results that were only in the 55% to 60% range. However, depending on the nature of the intervention/guidance that is being provided, it can be envisioned that if the system is unsure about the user's classification, some minimal adjustments can be done, followed by continued tracking to see if performance improves. Nevertheless, further research should be conducted to improve the presented accuracies. On the one hand, we envision that the addition of sequence features (e.g., scan path patterns) could provide even more information about the various tasks and user characteristics. On the other hand, eye-tracking data could be integrated with other sources, such as interaction data, if this information is available. Similarly, there are further sources of information that could potentially be added to such a system. For instance, it may be possible to infer the user's task through automatic graph analysis (e.g., based on computer vision techniques [Elzer et al. 2011] or natural language processing (e.g., by processing a visualization's caption).

We also obtained very interesting results regarding the more fine-grained details of each classification experiment. In particular, we found that depending on the goal of the classification, different features are most informative for different task/user characteristics (Q2). For example, we found that the legend usage increases for more complex tasks (i.e., tasks that have more data series) and label usage for generally more difficult tasks, suggesting that users could benefit from interventions relating to these particular AOIs. Similarly, we found that low perceptual speed users spend more time in the legend, suggesting that these users may benefit from interventions that particularly relate to this AOI (e.g., giving these elements more emphasis or providing easier access). These detailed analyses thereby not only provide evidence to what extent different characteristics can be inferred but also how a system may adapt to individual differences.

In terms of general trends regarding the most informative features, we found that for each of the classification runs, AOI-related features were crucial toward more accurate predictions. Thus, it may be argued that to build effective adaptive visualizations, a system needs to be aware of the currently active visualization. We therefore also showed that even in the case of the system not knowing this information a priori (e.g., if the adaptive component is not directly attached to the visualization system), it is possible to infer the visualization type solely based on a user's eye gaze with 70% accuracy. Again, this accuracy may potentially be improved with additional, more sophisticated features such as sequential scan paths.

Our experiments have only investigated two simple visualization techniques; however, there are many results that may be generalized to a wider array of visualization designs. In particular, we have shown that many of the important features are actually based on generic AOIs that are common to most types of visualizations, such as a graph's labels or legend. Similarly, while the study only focused on an artificial dataset involving student grades, the actual tasks were derived from an established set of general, low-level analysis tasks for information visualization [Amar et al. 2005] and may therefore be generalized to other application domains.

Our analysis also included a novel way of investigating task type/subtype similarity, since our study of confusion matrices (see Section 5.2) revealed common eye gaze patterns for certain types of tasks. This type of analysis may also be used for further research purposes, such as to determine which type of adaptation to provide to best support different task types/subtypes or common user strategies.

Last, although our experiments have shown results regarding the classification of different task and user characteristics—for instance, what to adapt to and to a certain extent how to adapt—more work needs to be carried out in terms of predicting when adaptive assistance is required. In particular, further research is necessary relating to the identification of potential user confusion or cognitive overload, which is related to the detection of "suboptimal usage patterns" that was discussed in related work by Gotz and Wen [2009].

6. CONCLUSIONS AND FUTURE WORK

In conclusion, we have presented research results showing that a user's eye gaze is a valuable source to infer a number of task and user characteristics. In particular, we have shown encouraging results using simple machine learning techniques on simple eye-tracking metrics, even after only partial data has been observed. The study has therefore provided a first step toward our long-term goal of designing user-adaptive information visualizations.

The next step of this research is to design user studies that focus on the effect of different adaptive interventions (e.g., highlighting, drawing reference lines, recommending alternative visualizations) on a user's performance, both in general and in relation to different tasks and individual user differences. These studies will also need to focus on different degrees of intervention intrusiveness, such as comparing fully adaptive versus mixed-initiative approaches. Following this investigation, we hope to develop a fully integrated adaptive information visualization system that is able to dynamically provide adaptive interventions that are informed by real-time user behavior data. Last, we will investigate the detection of user confusion/cognitive overload, as well as the usage of more complex features such as sequential scan paths, to improve on the results presented in this article.

REFERENCES

R. Amar, J. Eagan, and J. Stasko. 2005. Low-level components of analytic activity in information visualization. In *Proceedings of the 2005 IEEE Symposium on Information Visualization.* 15–21.

D. Bondareva, C. Conati, R. Feyzi-Behnagh, J. M. Harley, R. Azevedo, and F. Bouchet. 2013. Inferring learning from gaze data during interaction with an environment to support self-regulated learning. In *Artificial Intelligence in Education*. Lecture Notes in Computer Science, Vol. 7926, 229–238.

G. Carenini, C. Conati, E. Hoque, B. Steichen, D. Toker, and J. T. Enns. 2014. Highlighting interventions and user differences: Informing adaptive information visualization support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. 1835–1844.

S. M. Casner. 1991. Task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics* 10, 111–151.

C. Chen and M. Czerwinski. 1997. Spatial ability and visual navigation: An empirical study. *New Review of Hypermedia and Multimedia* 3, 67–89.

C. Conati and H. Maclaren. 2008. Exploring the role of individual differences in information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI'08)*. 199–206.

C. Conati and C. Merten. 2007. Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. *Knowledge-Based Systems* 20, 557–574.

F. Courtemanche, E. Aïmeur, A. Dufresne, M. Najjar, and F. Mpondo. 2011. Activity recognition using eye-gaze movements and traditional interactions. *Interacting with Computers* 23, 202–213.

J. P. Egan. 1975. *Signal Detection Theory and ROC-Analysis*. Academic Press.

S. Eivazi and R. Bednarik. 2011. Predicting problem-solving behavior and performance levels from visual attention data. In *Proceedings of the 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction (IUI'11)*. 9–16.

R. B. Ekstrom and U.S. Office of Naval Research. 1996. *Manual for Kit of Factor-Referenced Cognitive Tests*. Educational Testing Service.

S. Elzer, S. Carberry, and I. Zukerman. 2011. The automated understanding of simple bar charts. *Artificial Intelligence* 175, 526–555.

S. Few. 2005. *Keep Radar Graphs Below the Radar—Far Below*. Perceptual Edge.

K. Fukuda and E. K. Vogel. 2009. Human variation in overriding attentional capture. *Journal of Neuroscience* 29, 8726–8733.

J. Goldberg and J. Helfman. 2011. Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. *Information Visualization* 10, 182–195.

J. H. Goldberg and J. I. Helfman. 2010. Comparing information graphics: A critical look at eye tracking. In *Proceedings of the 3rd BELIV'10 Workshop: BEyond Time and Errors: Novel evaLuation Methods for Information Visualization (BELIV'10)*. 71–78.

D. Gotz and Z. Wen. 2009. Behavior-driven visualization recommendation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces* (*IUI'09)*. 315–324.

B. Grawemeyer. 2006. Evaluation of ERST: An external representation selection tutor. In *Proceedings of the 4th International Conference on Diagrammatic Representation and Inference (Diagrams'06)*. 154–167.

T. M. Green and B. Fisher. 2010. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In *Proceedings of the 2010 IEEE Symposium on Visual Analytics Science and Technology (VAST'10)*. 203–210.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter* 11, 10–18.

S. T. Iqbal and B. P. Bailey. 2004. Using eye gaze patterns to identify user tasks. *Presented at the the Grace Hopper Celebration of Women in Computing*.

A. Jameson. 2008. Adaptive interfaces and agents. In A. Sears and J. A. Jacko (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (2nd ed.). CRC Press, Boca Raton, FL, 433–458.

S. Kardan and C. Conati. 2012. Exploring gaze data for determining user learning with an interactive simulation. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization (UMAP'12)*. 126–138.

J. Mackinlay. 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics* 5, 110–141.

M. D. Plumlee and C. Ware. 2006. Zooming versus multiple window interfaces: Cognitive costs of visual comparisons. *ACM Transactions on Computer-Human Interaction* 13, 179–209.

F. J. Provost, T. Fawcett, and R. Kohavi. 1998. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th International Conference on Machine Learning (ICML'98)*. 445–453.

K. Rayner. 1995. Eye movements and cognitive processes in reading, visual search, and scene perception. *Studies in Visual Information Processing*, 3–22.

K. Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124, 372–422.

L. Sesma, A. Villanueva, and R. Cabeza. 2012. Evaluation of pupil center-eye corner vector for gaze estimation using a Web cam. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA'12)*. 217–220.

J. Simola, J. Salojärvi, and I. Kojo. 2008. Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research* 9, 237–251.

B. Steichen, H. Ashman, and V. Wade. 2012. A comparative survey of personalised information retrieval and adaptive hypermedia techniques. *Information Processing and Management* 48, 4, 698–724.

B. Steichen, G. Carenini, and C. Conati. 2013. User-adaptive information visualization: Using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI'13)*. 317–328.

D. Toker, C. Conati, G. Carenini, and M. Haraty. 2012. Towards adaptive information visualization: On the influence of user characteristics. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization (UMAP'12)*. 274–285.

D. Toker, C. Conati, B. Steichen, and G. Carenini. 2013. Individual user characteristics and information visualization: Connecting the dots through eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. 295–304.

D. Toker, B. Steichen, M. Gingerich, C. Conati, and G. Carenini. 2014. Towards facilitating user skill acquisition: Identifying untrained visualization users through eye tracking. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI'14).* 105–114.

M. L. Turner and R. W. Engle. 1989. Is working memory capacity task dependent? *Journal of Memory and Language* 28, 127–154.

M. C. Velez, D. Silver, and M. Tremaine. 2005. Understanding visualization through spatial ability differences. in: IEEE Visualization, 2005. VIS 05. In *Proceedings of IEEE Visualization (VIS'05)*. 511–518.

S. Westerman and T. Cribbin. 2000. Mapping semantic information in virtual space: Dimensions, variance and individual differences. *Journal of Human-Computer Studies* 53, 765–787.

C. Ziemkiewicz, R. J. Crouser, A. R. Yauilla, S. L. Su, W. Ribarsky, and R. Chang. 2011. How locus of control influences compatibility with visualization style. In *Proceedings of the 2011 IEEE Conference on Visual Analytics Science and Technology (VAST'11)*. 81–90.

Received August 2013; revised February 2014; accepted March 2014