User-Adaptive Information Visualization - Using Eye Gaze Data to Infer Visualization Tasks and User Cognitive Abilities

Ben Steichen, Giuseppe Carenini, Cristina Conati
Department of Computer Science,
University of British Columbia, Vancouver, Canada
{steichen, carenini, conati}@cs.ubc.ca

ABSTRACT
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 each individual user. To this end, this paper 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 the user's visualization tasks, as well as user cognitive abilities including perceptual speed, visual working memory, and verbal working memory. We show that such predictions are significantly better than a baseline classifier even during the early stages of visualization usage. These findings are discussed in view of designing visualization systems that can adapt to each individual user in real-time.

Author Keywords
Adaptive Information Visualization; Eye-tracking; Adaptation; Machine Learning.

ACM Classification Keywords
H.5.m.

INTRODUCTION
Information Visualization is a thriving area of Human-Computer Interaction that aims to help users in managing and understanding increasing amounts of information. While 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. In order to better assist each individual 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 of user-adaptive visualizations have already provided evidence for improved user performance (e.g. time on task, task accuracy), for instance by using click behavior to infer and adapt to suboptimal usage patterns [14], or by using a user's visualization selections to infer and adapt to a user's visualization expertise and preferences [15]. 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 such research on user-adaptive visualization in a number of aspects. First of all, we aim to expand the set of adaptation characteristics towards general (low-level) visualization task types, task complexity, as well as users' cognitive abilities (other than expertise) that have been shown to influence visualization performance. Secondly, while 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 using additional (and potentially complementary) non-interactive data sources (e.g. eye tracking). Thirdly, while 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 paper, we address the first two aspects by investigating to what extent a variety of visualization tasks and three different cognitive abilities (perceptual speed, visual working memory and verbal working memory) can be inferred from a user's eye gaze behavior. We focus on gaze behavior because visual scanning and processing are fundamental components of working with any visualization (and the only components for non-interactive visualizations). Specifically, we ask the following two research questions:

Q1. To what extent can a user's current task and/or long-term cognitive abilities be inferred from eye gaze data?

Q2. Which gaze features are the most informative?

The motivation of this particular work is two-fold. First of all, in order to provide appropriate adaptive support, an adaptive system needs to know about the user's *current*

IUI'13, March 19–22, 2013, Santa Monica, CA, USA.

task. For example, if the system knows that the user is currently trying to "filter" based on a particular data series, the system could adaptively deemphasize non-relevant data to reduce the user's cognitive load. Similarly, if a system knew a user's *cognitive abilities*, appropriate interventions could be provided to better assist this user. For example, since low perceptual speed has already been shown to lead to lower performance (in terms of speed and accuracy) [3][31][29], such users would benefit most from adaptive interventions.

Secondly, by analyzing which features are most informative, we may get an initial insight into what type of interventions would be most suitable. For example, as will be shown in the paper, different Areas of Interest (AOI) of a visualization provide considerable information about different task/user characteristics. Such results suggest that adaptations that are particularly tailored towards these AOIs may be most effective in order to support different tasks and/or user characteristics.

The remainder of this paper is structured as follows. First, we provide an overview of related research in adaptive visualization, 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 above. Finally, we conclude with a discussion of the overall findings and outline several directions for future work.

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 [27], adaptive hypermedia [27], desktop assistance [19], as well as e-learning [19]. 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* [2][22], rather than dynamically inferring individual properties *during* visualization usage. An exception to this non-adaptive paradigm is presented in [15], where users' visualization expertise and preferences are dynamically inferred through monitoring visualization selections (e.g. how long it takes a user to decide on 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 [15] 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 [14] actively monitors real-time user behavior during visualization usage in order to infer needs for intervention. In their work, interaction data (i.e. mouse clicks) are constantly tracked in order to detect suboptimal usage patterns, i.e. activities of users that are of a repetitive (hence inefficient) nature. Each of these sub-optimal patterns indicates that an alternative visualization may be more suitable to the current user activity. The patterns used in their paper include *scanning* (i.e. a user is iteratively inspecting over similar visual objects), *flipping* (iteratively changing filter constraints), *swapping* (repeatedly re-arranging 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 [15], namely they recommend alternative visualizations that may be more suitable for the current activity (for example, 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. Secondly, 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. Thirdly, the patterns do not try to infer general (low-level) visualization tasks (e.g. *filter*, *compute derived value*). Lastly, their approach does not attempt to adapt to any individual user characteristics.

As mentioned before, since visual scanning and processing are fundamental components of working with *any* visualization (they are in fact the only components for non-interactive visualizations), it is important to consider eye-tracking as a source of real-time information on user behavior. Although such 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 webcams) will enable the widespread application of such techniques in the near future [25]. 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 [21]. Similarly, research in this field has investigated the impact of individual user differences on basic reading and search tasks [24]. 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. Such research has typically focused on either identifying pattern differences for different visualizations

[13] or task types (e.g. reading vs. mathematical reasoning) [18], or on explaining differences in user accuracy between alternative interfaces [23]. However, such 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, such analyses have typically consisted of offline processes that require further human analysis (e.g. manually analyzing eye gaze coordinate plots [18]). In terms of actually using raw eye tracking data for real-time prediction, most research has so far focused on identifying cognitive processes while performing non-visualization activities, for example during exploratory e-learning [20][4], quizzes [5], simple puzzle games [7], or information search tasks (e.g. word search) [26]. 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 above approaches have 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, Ziemkiewicz et al. [32], as well as Green and Fisher [16] have looked at the influence of a user's personality traits, showing that locus of control (internal vs. external) can impact visualization performance. Similarly, cognitive measures such as perceptual speed and visual working memory have been shown to influence a user's ability to complete a task effectively [3][31]. For example, it has been shown that users with high perceptual speed had significantly faster completion times and accuracy on certain tasks. These results have been confirmed and extended in a recent study by Toker et al. [29], where perceptual speed, working memory, and user expertise were shown to influence not only a user's task performance, but also satisfaction regarding different visualization types. More specifically, they showed that depending on the respective cognitive abilities, users performed differently and preferred different visualizations. Most recently, it was also found that such individual user differences have an impact on different user eye gaze measures [28], which directly serves as the motivation for the work in this paper on using gaze data to dynamically identify and adapt to user cognitive abilities.

USER STUDY

To reiterate, this paper 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, as well as the real-time detection of task and user characteristics in order to be able provide appropriate interventions (note that this paper only focuses on the latter part). For these purposes, we designed and ran a user eye tracking study with two basic visualization techniques, namely bar graphs (Figure 1, top) and radar graphs (bottom). By choosing two different types of visualizations, we aimed to investigate the generalizability of our results.

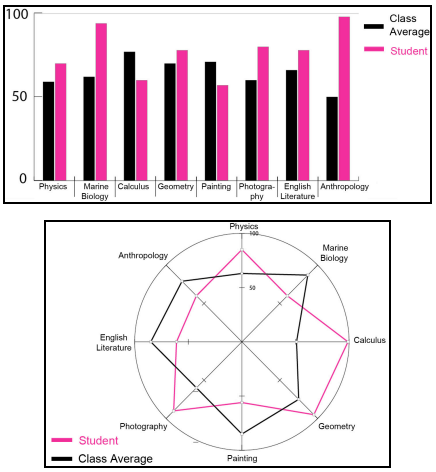

Figure 1. Sample bar (top) and radar graph (bottom)

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 [10], they are widely used for multivariate data. Furthermore, there are indications that radar graphs may be just as effective as bar graphs for more complex tasks [29].

Study Tasks

As part of the study design, 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. to be "representative of the kinds of specific questions that a person may ask when working with a data set" [1]. In particular, out of this taxonomy we chose the following five types: *retrieve value (RV), filter (FI), compute derived value (CDV), find extremum (FE),* and *sort (SO)*. Each of the tasks required users to evaluate student performance values in eight different academic courses (using an artificial data set). The tasks were chosen so that each of our two target visualizations would be suitable to support them. In order 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 8 academic courses (e.g., "In how many courses is Maria 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 5 single tasks, one for each task type (i.e. RV1, FI1, CDV1, FE1, SO1), and 4 double tasks (RV2, CDV2, FI2a, FI2b), meaning that the most fine-grained task type/complexity classification could consist of 9 classes (see classification experiments).

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. [31], as well as among the set that Conati and Maclaren [3] 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).

Study Procedure

Thirty-five subjects (18 female) participated in the experiment, ranging in age from 19 to 35. 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 CS students, while 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 one hour. Participants began by completing tests for three cognitive measures[1]: a computer-based OSPAN test for Verbal Working Memory [30] (lasting between 7 and 12 minutes), a computer-based test for Visual Working Memory [11] (10 minutes long), and a paper-based P-3 test for Perceptual Speed [8] (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 undergoing a calibration phase for the eye-tracker, participants performed 14 tasks per visualization (2x5 *single* and 1x4 *double*). The presentation order with respect to visualization type was fully counterbalanced across subjects. For each task, users were presented with a radar/bar graph displaying the relevant data, along with a textual question. Participants would then select their answer from a drop-down list and click OK to advance to the next task. The experimental software was fully automated and coded in Python.

Eye tracking measures & 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 upon basic eye-tracking measures (see Table 1).

Basic gaze measures	**Description**
Fixation rate	Rate of eye fixations per milliseconds
Number of Fixations	Number of eye fixations detected during an interval of interest
Fixation Duration	Time duration of an individual fixation
Saccade Length	Distance between the two fixations delimiting the saccade (d in Fig. 2)
Relative Saccade Angles	The angle between the two consecutive saccades (e.g., angle y in Fig. 2)
Absolute Saccade Angles	The angle between a saccade and the horizontal (e.g., angle x in Fig. 2)

Table 1. Description of basic eye tracking measures.

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 Fig. 2.), *Relative Saccades Angle* (e.g., angle y in Fig. 2) and *Absolute Saccade Angle* (e.g., angle x in Fig. 2.), as suggested in [12], because these measures are potentially useful for summarizing trends in user attention patterns within a specific interaction window, e.g., if the user's gaze follows a planned sequence (as opposed to being scattered).

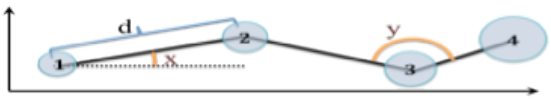

Fig. 2. Saccade based eye measures

In order to extract individual eye tracking features, the raw gaze data from the Tobii eye tracker was processed using customized scripts[2] written in the Python programming language. The scripts compute statistics such as sum, average and standard deviation over the eye tracking measures with respect to: (i) each overall task in the study, to get a sense of the complete interaction with the task (*task-level measures* from now on) and (ii) specific areas of interest (AOI), 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 in order to capture the distinctive and typical components of the two visualizations used in the study. Figures 3 and 4 show how these AOI map onto bar and radar graph components respectively.

[1] Note that these tests measure long-term cognitive traits, as opposed to short-term cognitive load. Therefore these tests do not have to be run during/in between tasks.

[2] These scripts are currently being generalized and will soon be released as an open-source toolkit named EMDAT (Eye Movement Data Analysis Toolkit).

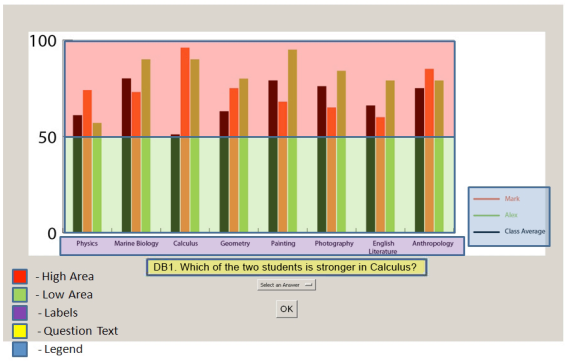

Figure 3: The five AOI regions defined over a bar graph

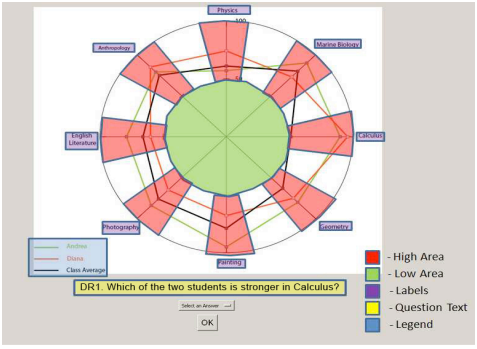

Figure 4: The five AOI regions defined over a radar graph

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 3); for the radar graph, it corresponds to the combined area of the 8 trapezoidal regions covering the data points (see Figure 4).
Low Area: covers the lower half of the data elements for each visualization.
Labels: covers all 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. Overall, a total of 79 features were calculated from the gaze data (see Table 2). For experimental purposes, we differentiated between a feature set that contained all features, including task-level and AOI features (called the *Full* set from now on), and one that did not contain features relating to AOIs (called the *No AOI* set), i.e. only containing the task-level features. This differentiation was chosen in order to evaluate the relative information gain attained when AOI features are available to the system.

TASK-LEVEL FEATURES
Fixations: *Total Number of Fixations, Fixation rate*
Fixation Durations: *Sum, Mean, Std. deviation*
Saccade Length: *Sum, Mean, Std. deviation*
Relative Saccade Angles: *Sum, Mean, Std. deviation*
Absolute Saccade Angles: *Sum, Mean, Std. deviation*
AOI-LEVEL FEATURES (for each AOI)
Total number of fixations in AOI
Sum & Mean of fixation durations in AOI
Time to first fixation in AOI
Time to last fixation in AOI
Longest fixation in AOI
Proportion of Total Number of Fixations in AOI
Proportion of Total Fixation Durations in AOI
Prop. number of Transitions From this AOI to every other AOI (5 separate measures)

Table 2. Eye Tracking Features

CLASSIFICATION EXPERIMENTS

The classification experiments described in this section use the above-mentioned features to infer a number of task and user characteristics. In particular, we investigate the extent to which these characteristics can be inferred from gaze data (Q1), as well as what gaze features are most important for classification (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) and *task complexity*, as well as 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.

Experimental process

Given the gaze features described in the previous section, we generated a number of datasets to simulate partial observation of gaze data during each task, e.g., only the first 10%, 20%, 30%, etc. These datasets were generated to simulate classification accuracies *while* a user is still interacting with (i.e., looking at) the visualization. The goal of this analysis is to investigate the feasibility of real-time interventions when integrating the classification component into a live user-adaptive visualization system.

We used the WEKA data mining toolkit [17] for model learning, as well as evaluation. For model learning, we tried a number of different classifier types (Decision Tree based, Support Vector Machine, Multilayer Perceptron, and Logistic Regression) with feature-selection and 10-fold

cross-validation for model evaluation. In all our experiments, Logistic Regression (LR from now on) was the classifier with the highest accuracy. In the following sections, the performance of this classifier is evaluated on both the *Full* and *No AOI* data sets. As a baseline for comparison, we use a classifier that always selects the most likely class (thus failing in all cases of the other classes), e.g., 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 the results of feature selection in detail. Instead, we discuss the impact of features only with respect to the performance of the *Full* vs. *No AOI* datasets.

Classification results for task characteristics

As explained in the *study procedure* section, users performed tasks of varying *type* and *complexity*. In this section we present classification results when tasks are defined at different levels of granularity.

Task type – 9 tasks

The most fine-grained analysis splits tasks into 9 different classes, one for each separate question type-complexity combination used in the study, e.g. *single retrieve value (RV1), double retrieve value (RV2)*, etc. Because of the high number of classes, this case represents a difficult multiclass classification challenge, with a baseline classification accuracy of only 15.45%. Nonetheless, the LR classifier using the full feature data set (LR-Full from now on), reaches a classification accuracy of 56.60% after seeing all the available data (see Figure 5), and starts having significantly higher accuracy than the baseline after seeing only 10% of the data.

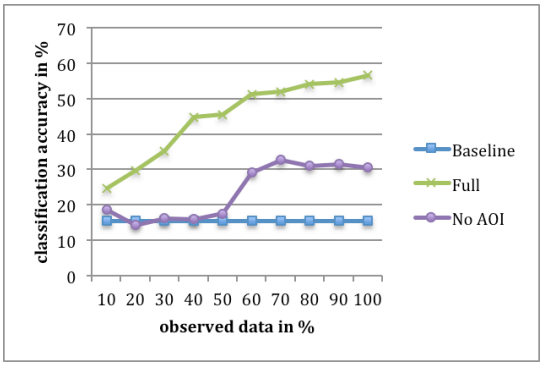

Figure 5. Task type – classification accuracy (9 different tasks)

Moreover, classification accuracy grows continuously as more gaze data becomes available, going over 50% after seeing 60% of the data (as shown in Figure 5). The average classification accuracy over time for LR-Full is 44.81%. 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 23.80% and its maximum accuracy after seeing all the data is 30.61%, both statistically significantly lower than the corresponding accuracies for LR-Full. Moreover, the accuracy of 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* vs. *No AOI* data sets indicate that AOI-related features have a strong impact on classification accuracy for task type at this granularity.

Task type- 5 tasks

In addition to the fine-grained task analysis involving 9 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, e.g., ignoring the difference between *retrieve value* when one student is involved (RV1) as opposed to when two students are involved (RV2). Ignoring this difference leaves us with 5 different classes, corresponding to 5 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 5-class classification task is very meaningful, because the classes represent general tasks types recognized as being common in information visualization.

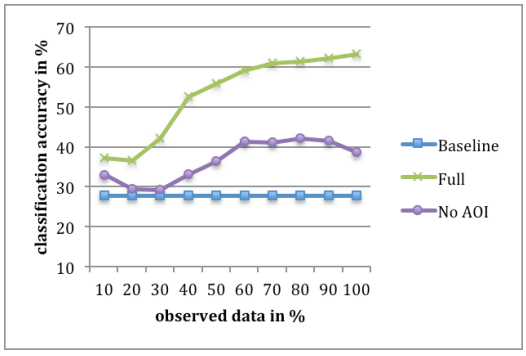

Figure 6. Task type – classification accuracy (5 tasks)

For this experiment, LR-Full reaches an average accuracy of 53.18% over time, an accuracy of 55.84% after seeing 50% of the data, and a maximum accuracy of 63.32% after seeing all the data. LR-Full statistically significantly outperforms the baseline's accuracy (27.863%) from the start. As was the case with 9 tasks, removing AOI-related features statistically significantly reduces accuracy, as shown by the performance of LR-NoAOI in Figure 6. This classifier reaches an average accuracy of 36.61% over time, an accuracy of 36.34% after seeing 50% of the data, and a

maximum accuracy of 42.25% after seeing 80% of the data. Moreover, this classifier only starts to be statistically significantly better than baseline after seeing 40%.

When analyzing sources of errors in the confusion matrix, we found two pairs of tasks that are most often confused with each 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., the *filter* task "… is used as a subtask in many other questions". Adaptations that particularly support this *filter* 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 Amar et al.'s taxonomy that "*sorting is generally a substrate for extreme value finding"* [1].

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 retrieve value (RV).

Task type - 3 tasks

When considering only 3 different task types, LR-Full reaches an average accuracy of 68.42% over time, an accuracy of 70.55% after seeing 50% of the data, and a maximum accuracy of 76.24% after seeing all the data. LR-Full statistically significantly outperforms the baseline's accuracy (48.14%) from the start.

As was the case with 9 and 5 tasks, removing AOI-related features statistically significantly reduces accuracy, as shown by the performance of LR-NoAOI in Figure 7. This classifier only reaches an average accuracy of 50.71% over time, an accuracy of 51.20% after seeing 50% of the data, and a maximum accuracy of 54.12 after seeing 70% of the data. Moreover, this classifier is only statistically significantly better than baseline between 50-70% of the data observations.

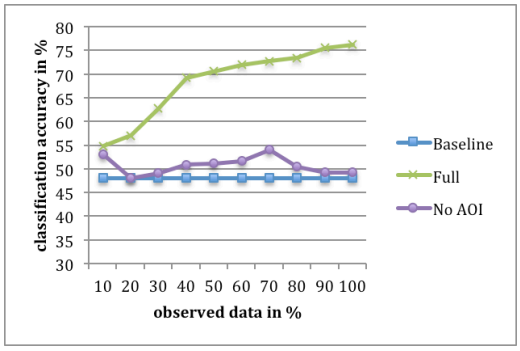

Figure 7. Task type – classification accuracy (3 tasks)

Task types – summary of results

In summary, we found that, across all task type granularities, LR with the full data set outperformed both the baseline and LR with the *No AOI* data set, showing the importance of having AOI-related features for task-type classification. Figure 8 summarizes the results in terms of average accuracy over time. As expected, accuracy for all 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 section), it is likely that we can increase accuracy on all our classification tasks.

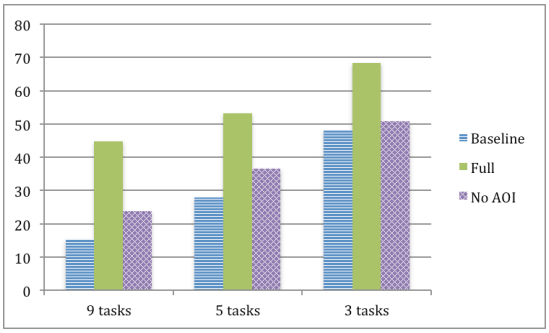

Figure 8. Task type – average classification accuracy over time for different task granularities

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 user study section, this distinction provides a measure for *task complexity*. LR-Full is still the most accurate classifier, with statistically significantly higher average accuracy (80.39%) over time than both the baseline classifier (72.69%) and LR-NoAOI (74.76%). 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, both LR classifiers performed higher, with accuracies reaching up to 84.45% for the full data set (see Figure 9). Accuracy again improved with more data being observed, and each of the feature sets outperformed the baseline after relatively low amounts of observed data (LR-Full is statistically significantly higher from the outset; LR-NoAOI after 40% of the data has been observed).

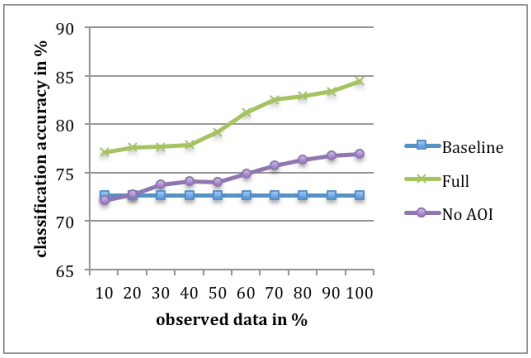

Figure 9. Task complexity – classification accuracy

Since task complexity consists of a simple two-class classification, it is also interesting to look into the more detailed feature selections and coefficients in order to analyze which exact features were contributing most to the classification (note that for multiclass logistic regression involving n classes, this analysis would involve analyzing n-1 feature selection results). Due to space limitations, we will only discuss the strongest features for LR-Full. With increased task complexity, we found that the use of the graph *legend* increased considerably, both in terms of proportionate amount of time spent (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 will in turn allow a user-adaptive system to provide adaptations for more complex tasks (e.g. provide support for better legend access and processing).

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 addition to reporting the classification accuracy, we will also discuss the gaze features that have the highest impact on classification.

In general, we found similar results across the three classification experiments. First of all, we found that LR-Full statistically significantly outperformed both the baseline and LR-NoAOI, with average accuracies for LR-Full ranging between 56-60%. While these average accuracies were rather low, it has to be noted again that these experiments are solely based on simple eye tracking measures, which may be improved using additional sources of information (see overall result discussion).

Several interesting observations were made when analyzing the accuracies at different data cut-off points. In particular, for each of the experiments, the peak accuracy of LR-Full was actually found after 20-40% of the data had been observed, as opposed to after all the data had been observed (as found in previous experiments). 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 (clearly presented in Figures 10, 11, 12), and that these patterns are increasingly "diluted" by other factors (e.g. task type) as the task goes on.

For *visual working memory*, the peak accuracy of 58.92% occurred after 40% of the data had been observed (see Figure 10). When analyzing the features that received the highest coefficient during feature selection for LR-Full, 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.

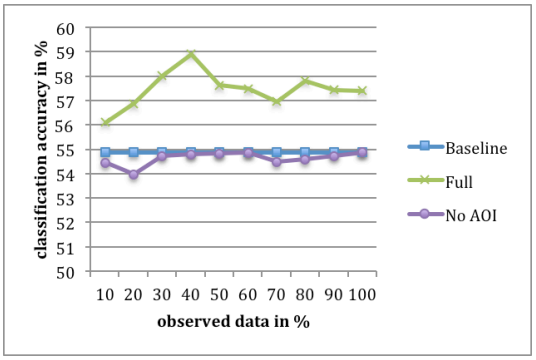

Figure 10. Visual Working Memory – classification accuracy

Similarly, for *verbal working memory*, the highest classification accuracy for both LR-Full (63.86%) and LR-NoAOI (61.09%) was found after observing only 20% of the data (see Figure 11). 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 20% of the data had been observed.

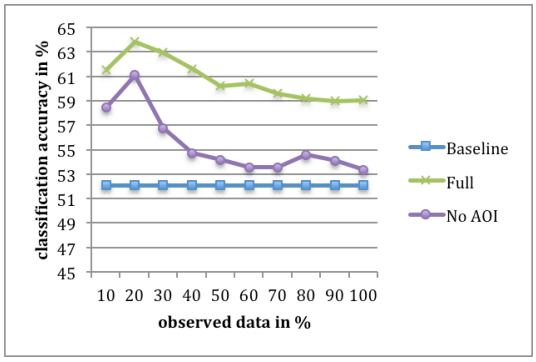

Figure 11. Verbal Working Memory – classification accuracy

A similar pattern was observed for the perceptual speed classification experiment, where the highest accuracy for LR-Full (59.84%) was found after only 20% of data had been observed (see Figure 12). When analyzing the feature selection results, 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. This finding may indicate that low perceptual speed users would benefit from adaptations relating to this particular AOI (e.g. through highlighting, facilitating easier access, etc.). In addition, high perceptual speed users had a shorter *longest fixation* and a higher *fixation rate*.

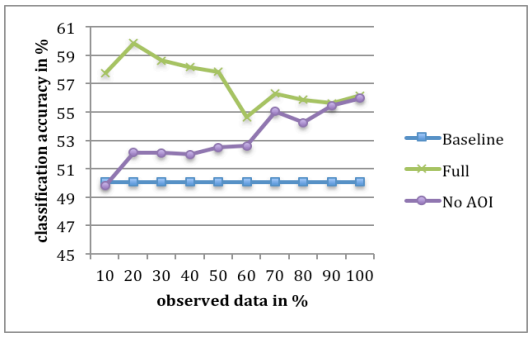

Figure 12. Perceptual Speed – classification accuracy

Visualization type

As shown in the classification results above, the inclusion of AOI-related features is critical towards generating predictions for both task types and cognitive measures. Having such AOI-related features requires knowing which visualization is currently active. While there are 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 (e.g., 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 statistical analysis tools, etc.), it would be first necessary to infer the currently active visualization type in order 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 is not applicable in a realistic scenario. For the LR-NoAOI classifier, the average accuracy is 64.56%, which is statistically significantly higher than the baseline (52.69%). As shown in Figure 13 (note that LR-Full is included for completeness), the accuracy of LR-NoAOI continuously grows as more gaze data is observed, reaching 66.34% after 50% of the data has been observed, and a maximum of 70.26% after all data has been observed. All accuracies are statistically significantly higher than baseline after only 20% of the data has been observed. While these results are encouraging, further research needs to be conducted in terms of improving accuracies in order to employ such techniques in a live system (see discussion).

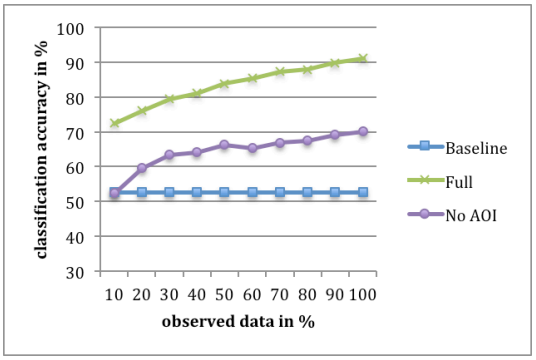

Figure 13. Visualization type – classification accuracy

Regarding the feature selection for this NoAOI classifier, we found that users have different viewing patterns in terms of *path angles*. Specifically, users have more horizontal viewing patterns in the bar graph (lower *mean absolute 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 *std. dev. of absolute path angles*).

Summary of Results & 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 and/or 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 findings regarding these research questions.

We found that a user's eye gaze behavior provides evidence about each of the visualization tasks and cognitive abilities. In particular, we show that for each classification task, gaze-behavior-based predictions outperform a baseline classifier (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.

It may be argued that from a practical point of view, the accuracies are not yet high enough to be exploited in a live system. In particular, the accuracies relating to the cognitive abilities yielded results that were only in the 55-60% 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 in order 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 task and user characteristics. On the other hand, eye tracking data could be integrated with other sources, for example interaction data if such information is available. Similarly, there are further sources of information that could potentially be added to such a system, for example it may be possible to infer the user's task through automatic graph analysis (e.g. based on computer vision techniques [9]) or natural language processing (e.g. by processing a visualization's caption).

We 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 label usage increases for more complex tasks, suggesting that users could benefit from interventions relating to this particular AOI. Similarly, we found that for low perceptual speed, users spend more time in the legend, suggesting such users may benefit from interventions that particularly relate to this AOI (e.g. giving such elements more emphasis or providing easier access). Such 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 towards more accurate predictions. It may therefore be argued that in order 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), it is possible to infer this 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.

While our experiments have only investigated two simple visualization techniques, 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 data set involving student grades, the actual tasks were derived from an established set of general, low-level analysis tasks for information visualization [1] and may therefore be generalized to other application domains.

Lastly, while our experiments have shown results regarding the classification of different task and user characteristics, i.e. *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 "sub-optimal usage patterns" that was discussed in related work by Gotz and Wen [14]. Similarly, if a system were able to detect "how well" a user is currently doing (i.e. if the system could infer user performance), adaptive assistance could be provided in cases of "bad" eye gaze patterns.

CONCLUSIONS & 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 towards 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, for example comparing fully-adaptive vs. mixed-initiative approaches. Following this investigation, we hope to develop a fully integrated adaptive information visualization system, which is able to dynamically provide adaptive interventions that are informed by real-time user behavior data. Lastly, we will investigate the detection of user performance and/or confusion, and we will investigate the usage of more complex features such as sequential scan paths to improve on the results presented in this paper.

REFERENCES

1. Amar, R.A., Eagan J., & Stasko, J.T. Low-Level Components of Analytic Activity in Information Visualization. In 16th IEEE Visualization Conference, (2005), 15-21.
2. Casner, S. M. A task-analytic approach to the automated design of graphic presentations. ACM Trans. on Graph., 10(2), (1991), 111-115.
3. Conati, C., & Maclaren, H. Exploring the Role of Individual Differences in Information Visualization. In Proceedings of the working conference on Advanced visual interfaces (AVI '08). ACM, New York, NY, USA, (2008), 199-20.
4. Conati, C., & Merten, C. Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. Know.-Based Syst., 20(6), (2007), 557-574.
5. Courtemanche, F., Aïmeur, E., Dufresne, A., Najjar, M., & Eboa, F.H.M. Activity recognition using eye-gaze movements and traditional interactions. In Proceedings of In-teracting with Computers, (2011), 202-213.
6. Dillon, A., & Shaap, D. Expertise and the perception of structure in discourse. Journal of the American Society for Information Science, 47(10), (1996), 786-788.
7. Eivazi, S., Bednarik, R. Predicting Problem-Solving Behavior and Performance Levels from Visual Attention Data. In the proceedings of 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI 2011, (2011), 9-16.
8. Ekstrom, R., French, J., Harman, H. and Dermen, D, Manual from Kit of Factor-References Cognitive Tests. Educational Testing Service (1976), Princeton, NJ.
9. Elzer, S., Carberry, S., & Zukerman, I. The automated understanding of simple bar charts. Artif. Intell. 175(2), (2011), 526-555.
10. Few, S. Keep Radar Graphs Below the Radar - Far Below: Information Management Magazine, (2005).
11. Fukuda, K., & Vogel, E.K. Human variation in overriding attentional capture. Journal of Neuroscience, (2009), 8726-8733.
12. Goldberg, J.H. & Helfman, J.I. Comparing Information Graphics: A Critical Look at Eye Tracking, BELIV ’10, (2010), 182-195.
13. Goldberg, J.H. & Helfman, J.I. Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. Information visualization 10(3), (2011), 182-195.
14. Gotz D., & Wen, Z.. Behavior-Driven Visualization Recommendation. ACM Int. Conf. on Intelligent User Interfaces, (2009), 315-324.
15. Grawemeyer, B. Evaluation of ERST – an external representation selection tutor. In Proc. of the 4th int. conf. on Diagrammatic Representation and Inference (Diagrams'06), (2006), 154-167.
16. Green, T. M. & Fisher, B. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In IEEE Visual Analytics Science and Technology (VAST), (2010), 203-210.
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. The WEKA Data Mining Software: An Update, SIGKDD Explor., 11 (1), (2009).
18. Iqbal, S.T., & Bailey, B.P. Using eye gaze patterns to identify user tasks. In: The Grace Hopper Celebration of Women in Computing, (2004).
19. Jameson, A. (2008). Adaptive Interfaces and Agents. The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications (2nd ed.), pp. 433–458.
20. Kardan, S., Conati, C.: Exploring Gaze Data for Determining User Learning with an Inter-active Simulation. Proc. 19th intl. conf. on user modeling, adaptation, and personalization, (2012), 126-138.
21. Keith, R. Eye movements and cognitive processes in reading, visual search, and scene perception. Eye Movement Research Mechanisms, Processes, and Applications. North-Holland Press, (1995).
22. Mackinlay, J. Automating the design of graphical presentations of relational information. ACM Trans. on Graph., 5(2), (1986), 110-141.
23. Plumlee, M.D., & Ware, C. Zooming versus multiple window interfaces: Cognitive costs of visual comparisons. ACM Trans. Comput.-Hum. Interact, 13, 2, (2006), 179-209.
24. Rayner, K. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124, (1998), 372-422.
25. Sesma, L., Villanueva, A., & Cabeza, R. 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), (2012), 217-220.
26. Simola, J., Salojarvi, J., & Kojo, I. Using hidden Markov model to uncover processing states from eye movements in information search tasks. Cognitive Systems Research. 9, (2008), 237-251.
27. Steichen, B., Ashman, H., & Wade, V. A comparative survey of Personalised Information Retrieval and Adaptive Hypermedia techniques. Information Processing & Management, 48 (4), (2012), 698–724.
28. Toker D., Conati C., Steichen B., & Carenini G. Individual User Characteristics and Information Visualization: Connecting the Dots through Eye Tracking. In Proc. of the ACM SIGCHI Conference on

Human Factors in Computing Systems, (CHI 2013), (2013), (to appear).

29. Toker, D., Conati, C., Carenini, G., & Haraty, M. Towards Adaptive Information Visualization: On the Influence of User Characteristics. In Proc. of the 19th international conference on user modeling, adaptation, and personalization, UMAP, (2012), 274-285.

30. Turner, M. L., & Engle, R.W. Is working memory capacity task dependent? Journal of Memory and Language, 28(2), (1989), 127-154.

31. Velez, M.C., Silver, D., & Tremaine, M. Understanding visualization through spatial ability differences. Proceedings of Visualization, (2005), 511-518.

32. Ziemkiewicz, C., et al. How Locus of Control Influences Compatibility with Visualization Style. Proc. IEEE VAST, (2011), 81-90.