Towards facilitating user skill acquisition - Identifying untrained visualization users through eye tracking

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

ABSTRACT
A key challenge for information visualization designers lies in developing systems that best support users in terms of their individual abilities, needs, and preferences. However, most visualizations require users to first gather a certain set of skills before they can efficiently process the displayed information. This paper presents a first step towards designing visualizations that provide personalized support in order to ease the so-called 'learning curve' during a user's skill acquisition phase. We present prediction models, trained on users' gaze data, that can identify if users are still in the skill acquisition phase or if they have gained the necessary abilities. The paper first reveals that users exhibit the learning curve even during the usage of simple information visualizations, and then shows that we can generate reasonably accurate predictions about a user's skill acquisition using solely their eye gaze behavior.

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

ACM Classification Keywords
H.5.m.

General Terms
Human Factors; Measurement.

INTRODUCTION

Individual user abilities, needs, and preferences have been shown to play an important role in the effectiveness of many human-computer interaction and information visualization systems. User differences can include medium- to long-term characteristics such as interests, personality, or cognitive abilities, as well as more short-term states such as cognitive load or affect. The benefits of dynamically adapting to these differences have already been demonstrated in a variety of human-computer interaction tasks and applications, such as menu based interfaces, web search, desktop assistance, or human learning [13]. One important characteristic that has received less attention (in terms of adaptation) is a user's experience or competence with a system, especially in the information visualization area. In particular, most visualizations typically require users to first acquire a certain set of skills, gained through practice, before they can efficiently process the displayed information

The long-term goal of the research presented in this paper is to devise visualizations that can ease the so-called 'learning curve' through adaptive support during a user's skill acquisition phase. Such support may include preventing untrained users from accessing advanced features, providing tooltips, offering tutorials, etc. As with any user-adaptive system design, the key challenges of this endeavor lie in (i) measuring the effect that a target user's characteristics (in our case *skills in using a given visualization*) have on user performance, (ii) detecting these characteristics in real-time, and (iii) providing adaptive help to best support the user's current needs and abilities.

In this paper, we focus on challenges (i) and (ii), namely on verifying the effect and supporting the detection of a user's skill acquisition with information visualizations. With respect to skill acquisition detection, we investigate the value of user eye gaze information as a data source, because visual scanning and processing are fundamental components of working with any information visualization system (and the only components for non-interactive visualizations). Our research questions are as follows:

1) To what extent can a user's skill level be predicted in real-time, using solely eye gaze data?

2) Which eye gaze features are most predictive?

3) To what extent is knowledge about the user's current visualization required for these predictions?

In order to answer these research questions, we leveraged data obtained from a study that involved users performing low-level visualization tasks with simple bar graphs. The paper shows that, despite our best efforts to control for any learning/ordering effects, participants indeed exhibited a learning curve even with these simple visualizations. The paper then shows that we can generate reasonably accurate predictions about a user's skill acquisition using classifiers that are trained solely based on eye gaze data. In particular, we show that from the outset (i.e., after only seeing a small part of a user's gaze data) our classifiers outperform a

IUI'14, February 24–27, 2014, Haifa, Israel.

ACM 978-1-4503-2184-6/14/02...$15.00.
http://dx.doi.org/10.1145/2557500.2557524

simple baseline, and that accuracies can reach up to 64%. To investigate the generalizability of our approach, we provide comparative results for classifiers that use *Visualization-Specific* features versus *Generic* feature sets, and we show that some the most predictive features are in fact visualization-independent.

RELATED WORK

Effects of Individual Differences on User Performance with Visualizations

Recent visualization research has shown that user characteristics can significantly influence user performance. There is substantial evidence that cognitive measures such as *perceptual speed*, *visual working memory* and *verbal working memory* influence user effectiveness and satisfaction when working with a visualization [25][27]. The personality trait known as *locus of control* (internal vs. external) has been shown to impact visualization performance (e.g., [29]). Several researchers have looked at the impact of *domain expertise* on performance with visualizations (i.e., a user's expertise in the task domain, as opposed to expertise with the visualization itself). For example, Dillon [9] discusses how domain expertise (e.g., experience reading academic journals in cognitive science) has been repeatedly shown to play a significant role in predicting performance with various visual navigation tools, and that this should be taken into account when designing visualization systems. Similarly, domain expertise consisting of measuring prior technical training (e.g., in statistics, psychology, etc.) has been shown to play a significant role in visualization performance, e.g., [16][19].

System Expertise in Adaptive Systems & Skill Acquisition

Outside information visualization, there has also been work on modeling and adapting to a user's *system expertise*, i.e. to the user's level of familiarity with the interactive system being used. For example, Bunt et al. [5] devised and evaluated MICA, a mixed-initiative GUI-customization tool that provided suggestions on how to personalize the menus of a word processor, by considering, among other factors, the user's expertise with the word processor. Expertise levels were defined based on how much time a user took to perform menu selections with the interface. However, the ability to track a user's system expertise in real time was not implemented. An evaluation of MICA, in which expertise was assessed via a pre-questionnaire, showed better performance with and higher preference for MICA compared to a version that provided the customization functionality without personalized suggestions. Linton & Shaefer [17] generated a model of expert usage of a word processor based on the frequency, sequence, and number of distinct menu commands displayed by the users of the application. This model was then used to generate recommendations on which functionalities to use for users who diverged from the expert model. In this paper, we contribute to this line of work by looking at system expertise with a visualization, and at whether we can track how it evolves during usage, a problem that to our knowledge has yet to be addressed in research of user-adaptive interaction.

In perceptual psychology, numerous theoretical models exist on this topic of expertise, or *skill acquisition* (see [1] for an overview). While it is not within the scope of this paper to argue for the correctness or fit of any of these theories, we focus on the fact that a typical method used in psychology for tracking how user performance improves with practice is by using a *learning curve* [23]. Learning curves are also frequently used in HCI to compare and evaluate the effectiveness of various systems, including information visualization systems (e.g., [28][22][20]). In this paper, we leverage the concept of learning curve as a way to identify two broad stages of a user's skill acquisition, which we then use to evaluate the detection of a user's skill using eye-gaze data.

Eye-tracking in User Modeling for Adaptive Systems

Several studies have examined the value of using eye tracking data as an input source for real-time modeling of relevant user characteristics. For example, Qu & Johnson [21] showed that user gaze behaviors can help predict users' motivation during interaction with an intelligent tutoring system. Kardan et al. [14] and Bondareva et al. [4] showed that eye tracking data can be used to predict student learning with two different educational environments, and that this prediction can be performed early enough to possibly provide adaptive interventions that can foster learning. D'Mello et al. [8] evaluated an intelligent tutoring system that both detected and reacted to students' lack of attention based on gaze patterns. They found that this gaze-reactive tutor had a positive impact on student learning. Bednarik et al. [3] used eye tracking features in order to predict users' problem-solving strategies (e.g., evaluation, intention, planning, etc.), as well as user performance while solving a visual puzzle. They examined the effect of window-size on feature extraction, and found that, in general, increased window sizes led to an improvement in classification. Conati & Merten [7] combined gaze data with information on user's actions to predict user meta-cognitive behavior (e.g., self-explanation) within an exploratory learning environment. Steichen et al. [24] showed very positive results in using gaze data to recognize in real time a user's tasks and cognitive traits (e.g., perceptual speed, verbal working memory) while interacting with two simple visualizations (bar and radar graphs). Similar to the work in [24], the data we use in this paper comes from a study involving users who perform low-level visualization tasks using simple bar graph visualizations. Here, however, we use gaze data to predict a *user's skill acquisition phase* in working with the visualization.

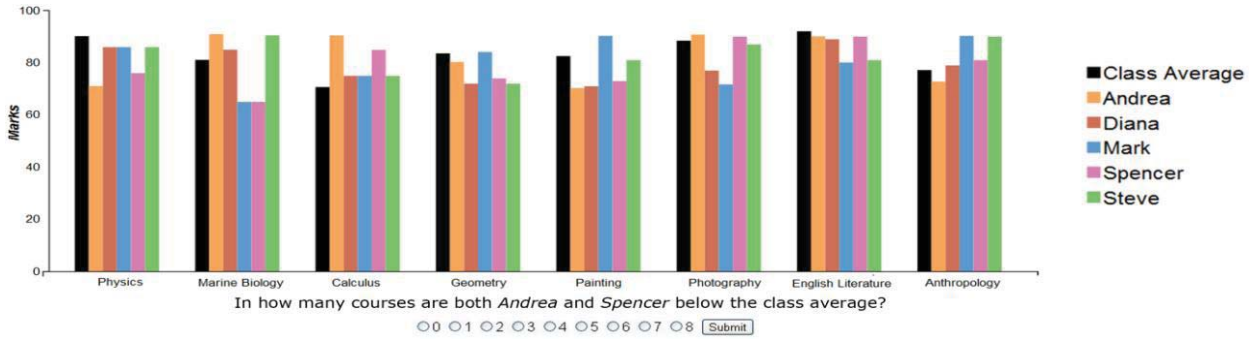

Figure 1. Example bar graph visualization as used in the experimental task

USER STUDY

In this section, we provide an overview of the study we conducted to gather empirical data (including eye tracking data) about bar graph processing. The primary purpose of this study was to investigate the relative effectiveness of several 'visual prompts' designed to help visualization processing, as well as the relative effect of different user traits and task complexity [6]. Here, we leverage the data from this study to investigate user skill acquisition with simple visualizations. In the next few sections, we provide a summary of the main components of the study sufficient for the purposes of this paper.

Experimental Visualizations and Tasks

In the study, participants were given bar graph visualizations, along with textual questions for them to answer, relating to the displayed data (see Figure 1). We selected bar graphs as the information visualization for this study because (i) they are a very common and basic visualization, and (ii) there is already research that shows that several types of individual differences can play a role in the effectiveness of bar graphs [25], suggesting that user-adaptive techniques could be of benefit. As mentioned above, some of the visualizations (fully randomized) contained one of four highlighting interventions (see Figure 2) designed to guide the user's focus to a specific subset of data within the bar graph that is relevant to answer the associated question. The experimental software was fully automated and ran in a web-browser, with the visualizations and interventions being programmed using the D3 visualization framework [7]. The experiment was conducted on an Intel Core i7, 3.4GHz, with 4GB of RAM, connected to a Tobii T120 eye-tracker as the main display.

The study tasks involved comparing individuals against a group average (data points in the bar graph) on a set of dimensions (data series in the bar graph). For variety, the task questions were drawn from four different domains. All tasks involved the same number of data points (six, including the average) and series (eight). Two types of tasks were chosen from a set of primitive data analysis tasks that Amar et al. [2] identifies as "largely capturing people's activities while employing information visualization". The first task type was Retrieve Value (a relatively simple task), which consisted of retrieving a specific individual in the target domain and comparing it against the group average; (e.g., "Is Michael's grade in *Chemistry* above the class average for that course?"). The second task type was Compute Derived Value (a more complex task type), which required users to first perform a set of comparisons, and then compute an aggregate of the comparison outcomes; (e.g., "In how many cities is the movie *Vampire Attack* above the average revenue and the movie *How to Date Your Friends* below it?").

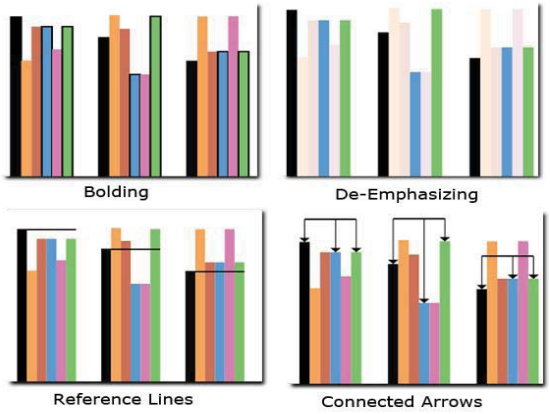

Figure 2. Example visualizations with added prompts

Study procedure

The study had 62 subjects, ranging in age from 18 to 42. Participants were mostly recruited via dedicated systems at our university, resulting in a variety of students from diverse backgrounds (e.g., Psychology, Forestry, Computer Science, Finance, Fine Art, German, Commerce). We also recruited 7 non-student participants such as a non-profit community connector, a 3D artist, and an air combat systems officer. The experiment was a within-subjects study, fitting in a single session lasting at most 90 minutes, with each participant completing a total of 80 trials covering combinations of task type and visual prompts.

Participants began by completing a number of pre-study questionnaires and cognitive tests (not used in this paper). Next, participants underwent a training phase to expose them to bar graphs, the study tasks, and the visual prompts. The training phase first involved familiarizing users with all

of the features of our visualization layout (e.g., x-axis, y-axis, legend mapping, labels, bars etc.), followed by a series of practice tasks that exposed users to the various task types as well as the layout of the interventions. Participants then underwent a calibration phase for the eye-tracker, before starting the study trials. Participants performed 40 of the 80 study trials, followed by a 5-minute break. After the break, the eye-tracker was re-calibrated and the participant performed the remaining 40 trials. The 80 trials were fully randomized in terms of experimental conditions (i.e., task complexity, interventions). Lastly, participants took a post-questionnaire designed to gauge their evaluations of each intervention's usefulness, as well as their relative preferences (not used in this paper).

EVIDENCE OF SKILL ACQUISITION: MAIN EFFECT OF TRIAL ORDER

A prior analysis of the study data (based on an ANOVA repeated measures) revealed interesting effects of interventions, task type, and user characteristics on performance (see [6]). However, despite our best efforts to control for any learning/ordering effects, (e.g., by training each user with the visualization system at the beginning of the study as well as by fully randomizing the experimental conditions), a General Linear Model repeated measures revealed a main effect of trial order on task completion time, ($F_{79,1142}$= 6.85, p < .001). This result indicates that users improved significantly over time, independently from the other experimental factors (e.g., task type, visual prompt type). Note that for task accuracy (measured for each trial as either correct/incorrect), a Friedman's ANOVA indicated no main effect of trial order on accuracy, ($\chi^2(62)$= 115.83, p = .262), likely due to a ceiling effect.

Figure 3 shows the learning curve for our study data, which plots the average performance across all users over the 80 study tasks, in order of completion (i.e., average user performance on the i^{th} trial, where i ranges from 1 to 80).

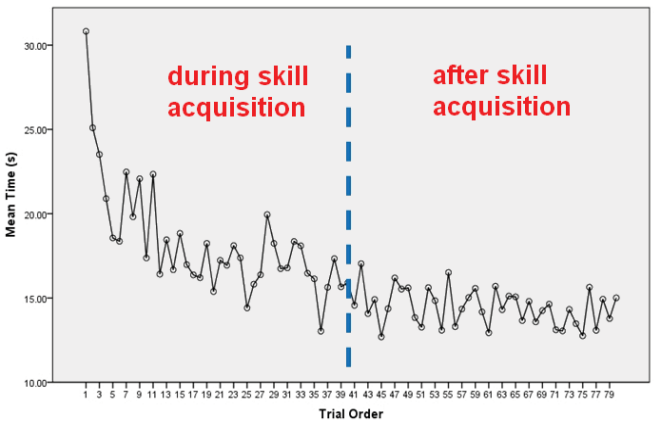

Figure 3. Learning curve showing performance improvement. The blue line separates trials into two general stages of skill acquisition, which we call *during* and *after*

The curve clearly indicates an improvement in performance, as characterized by a descending slope for roughly the first half of the trials (left of the dotted blue line in figure 5). In the second half of trials (right of the dotted blue line), however, performance appears to stabilize (as indicated by a reduced variance across trials). These results suggest that users were acquiring relevant skills during the initial part of the experiment, and that the skills are related to the processing of the visualizations, because the learning effects in Figure 3 are independent of both the type of task performed and the intervention received (recall that the actual task and intervention type seen for the i^{th} trial varied across users, because of randomization). The presence of this learning effect therefore suggests that it may be useful to track and facilitate a user's skill acquisition phase when working with a visualization. In the next section, we present the classifiers we built to detect a user's skill acquisition phase using a user's eye gaze data.

CLASSIFIERS FOR SKILL ACQUISITION DURING VISUALIZATION PROCESSING.

Classification Labels

Because participants were given a break at the halfway point in the study (after 40 trials), and much of the skill acquisition effects in Figure 3 appeared to happen during this first part of the study, we opted to use this break as the boundary to generate labels for classification of skill acquisition. The 40 trials before the break are hence labeled *during* (skill acquisition), and the 40 trials after the break are labeled *after* (skill acquisition). This choice is further supported by the fact that the difference in completion time between the two phases (i.e., *during* vs. *after*) is statistically significant (using an ANOVA, p < .001). The average completion time for tasks *during* skill acquisition was 18.2s (SD=10.7), whereas the average performance *after* skill acquisition was 14.4s (SD=8.2). Conceptually, trials labeled as *during* represent instances where (in general) a user is still practicing/undergoing skill acquisition with the visualization system, whereas trials *after* represent instances where a user has become practiced/competent. The benefit of labeling the trials in this manner, as opposed to defining skill acquisition in terms of some fixed value(s) of time, is that the labels are thus relative, meaning that this definition can be expanded to other visualization (and non-visualization) systems. Specifically, we envision labeling additional task interaction data sets taken from other visualization systems, where the concept of *during* skill acquisition and *after* skill acquisition can be transferred relative to the sets of tasks being performed for that given interface.

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 processed and analyzed to derive attention patterns. Following the approach in [14]

and [24], we generated a large set of eye-tracking features by calculating statistics upon basic eye-tracking measures (see Table 1 & Table 2).

Table 1. Basic Gaze Measures

Basic gaze measures	Description
Fixation Count	Count of number of fixations
Fixation Duration	Time duration of an individual fixation
Saccade Length	Distance between the two fixations delimiting the saccade (d in Figure 4)
Relative Saccade Angles	The angle between the two consecutive saccades (e.g., angle y in Figure 4)
Absolute Saccade Angles	The angle between a saccade and the horizontal (e.g., angle x in Figure 4)

Of these basic measures, *Fixation Count* and *Fixation Duration* are widely used in eye tracking studies. In addition, we included *Saccade Length* (e.g., distance d in Figure 4); *Relative Saccades Angle* (e.g., angle y in Figure 4); and *Absolute Saccade Angle* (e.g., angle x in Figure 4); as suggested in [10], 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).

The raw gaze data from the Tobii eye tracker was processed using our open-source data analysis toolkit, which is freely available for download and extension by the research community[1]. The toolkit computes features such as sum, average, and standard deviation over the eye tracking measures with respect to (i) the overall screen, to get a sense of the complete interaction with the task (*Overall Features* from now on) and (ii) specific areas of interest (AOI), identifying sub-parts of the interface that may be relevant for understanding a user's attention processes (*AOI-level Features* from now on). The total range of features computed by EMDAT is shown in Table 2.

Table 2. Features calculated based on basic gaze measures

Overall Features
Total Number of Fixations, Fixation rate
Sum, Mean, Std. deviation of Fixation Durations
Sum, Mean, Std. deviation of Saccade Length
Sum, Mean, Std. deviation of Relative Saccade Angles
Sum, Mean, Std. deviation of Absolute Saccade Angles
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
Number of Transitions From this AOI to every other AOI (*n*(n-1)* separate measures, where *n* is the number of AOIs)

[1] http://www.cs.ubc.ca/~skardan/EMDAT/

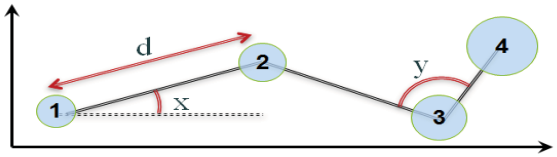

Figure 4. Saccade based eye measures

AOI definitions

Visualization-specific AOIs can be useful for gaining detailed insights on how certain visualization components are being processed by users (e.g., Toker et al. *[26]* found that users spend more time looking at the legend AOI during difficult tasks). Additionally, [4] and *[24]* have shown that by including visualization-specific AOIs, higher classification accuracy can be achieved in terms of predicting properties such as learning gain, cognitive abilities, visualization type, and task type. In our study, the *Visualization-Specific* set contained the following 6 AOIs:

- *High Area*: covers the upper half of the data elements of the visualization which corresponds to a rectangle over the top half of the vertical bars.
- *Low Area*: covers the lower half of the data elements.
- *Labels*: covers the data labels.
- *Question Text*: covers the text describing the task to be performed.
- *Legend*: covers the legend, which shows the mapping between the data series and the color of the visualization elements.
- *Answer Input:* covers the task response radio buttons and submit button.

Since one of the aims of this paper is to investigate how accurately we can predict user task acquisition independent of the visualization (research question 3), we also explore the possibility of making predictions using *Generic AOI* sets, which are not defined in terms of any particular visualization or interface. Specifically, we analyze if any of them may be suitable/adequate alternatives to the *Visualization-Specific* AOIs. In total we devised *5 Generic AOI* sets (referred to as *2x2*, *3x3*, *4x4*, *5x5*, and *X grid*), with each of them consisting of grid-like AOIs and differing only in terms of granularity/layout (see Figure 5).

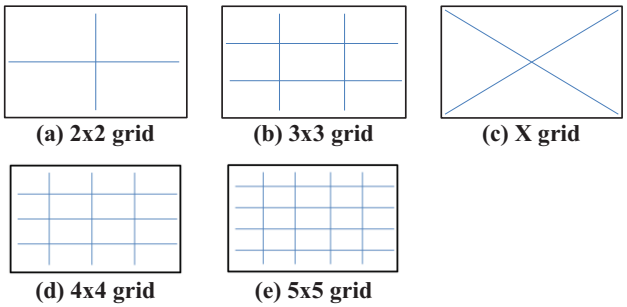

Figure 5. Generic AOI grids

Data sets and evaluation process

Similar to the analysis performed in [24], the goal of our classification experiments is to predict the correct labels (in our case *during* vs. *after*) for individual user trials based on eye tracking features. In order to explore the potential of classifying users while they are engaged in a task (to be able to then provide dynamic help), we generated a number of datasets that simulated partial observation of a user's eye gaze data. In the next section, we will first look at datasets generated based on *percentage* of the overall interaction during a trial (e.g., 10%, 20%, etc.), and then at datasets generated based on *absolute time* intervals (e.g., 1sec, 2sec, etc.).

Additionally, in order to study the importance of having knowledge regarding which specific visualization (e.g., bar graph) a user is currently engaged with, we compare the classification performance between each of the various AOI sets defined in the previous section, i.e., the *Visualization-Specific* set, the 5 *Generic* sets, as well as a classifier with no AOIs ('*None'*). As a baseline, we use a simple classifier that always selects the most likely class.

For each of our classification experiments, we used the WEKA data mining toolkit [11] for model learning and evaluation. In particular, we used a Logistic Regression classifier, both for its simplicity, and because it has previously been found to be the best performing classifier for experiments involving eye gaze data [4][24]. For all our experiments, classification accuracy is computed using 10-fold cross validation.

CLASSIFICATION RESULTS

Percentage-based Time Intervals

To evaluate how feasible it is to classify a user's skill acquisition phase from gaze data, we first generated datasets consisting of incremental percentages (time intervals) of interaction data, following an approach proposed in [4][14][24]. This approach is a good proof of concept to determine classification accuracy given different amount of interaction data, without having to worry about variances in users' completion times, e.g., it allows us to verify whether the first 10%, 20% etc., of a user's interaction are particularly good for the given classification task.

The results of these analyses are shown for each AOI set in Figure 6. The trends in the figure show that the *Visualization-Specific* classifier generally performs best, indicating that, not surprisingly, knowing which visualization the user is working with improves skill acquisition detection. A one-way ANOVA with AOI-type as the independent variable (8 levels), and the *average over time accuracy* as the dependent measure shows that there is indeed a statistically significant effect of AOI-type on classification accuracy ($F_{7,783}$= 162.7, p < .001). Bonferroni-adjusted pairwise comparisons further qualify these effects as follows:

- All classifiers are significantly better than the baseline classifier (p < .001);
- The *Visualization-Specific* classifier (M=62.7, SD=1.0) performs significantly better than all the *Generic AOI* classifiers (M=60.59, SD=0.7) and the *None* classifier (M=59.8, SD=0.6) (p < .001);
- The *3x3 AOI* classifier (M=61.1, SD=1.2) performs significantly better than *None* (M=59.8, SD=0.6) (p < .001);
- There are no other significant differences among the *None* and *Generic AOI* classifiers.

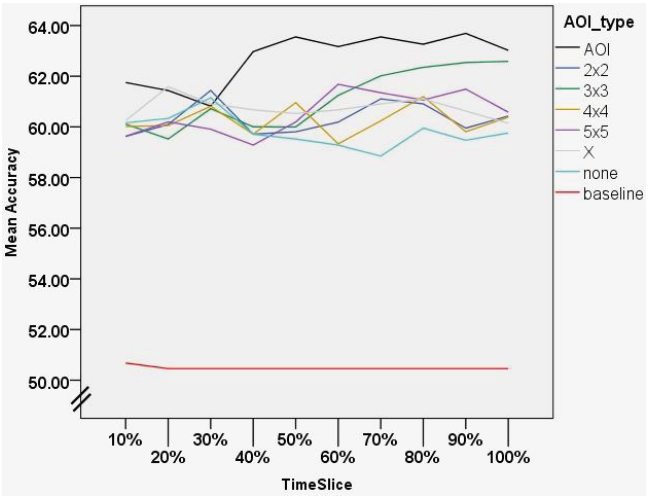

Figure 6. Classification accuracy of user skill acquisition state using percentage-based eye tracking features

Although our results indicate that visualization-specific gaze features can significantly improve the classification of skill acquisition, it should be noted that the 'specific' AOIs in our visualizations could easily be transferred to different visualizations, since they mainly consist of common visualization components (e.g., labels, legend). However, if the goal is to design visualization-independent classifiers (i.e., not just common visualizations), our results indicate that classifiers based on generic AOIs also perform reasonably well, considering that this is based purely on generic eye gaze features (which could potentially be combined with other sources such as interaction or mouse-tracking data). The results also suggest that the AOI granularity needs to strike a compromise between specificity and sensitivity. In particular, results indicate that using the 3x3 grid is the best generic AOI alternative given that it is the only set that performed significantly better than using no AOI information. This points to a possible tradeoff between having too few AOIs (i.e., not enough precision to track meaningful gaze movements across AOIs) versus too many AOIs (i.e., granularity being too fine-grained and gaze movements becoming too noisy across small AOIs).

Absolute Time Intervals

While the approach described in the previous section (i.e., investigating classification accuracy based on percentage of

available data) can give valuable insights into trends and patterns of classification accuracy, it requires a task to be fully completed in order to determine what constitutes 100% of the interaction. In practice, a real-time classifier needs to make predictions without knowledge of when the target task will be completed. For this reason, we also generated partial observation datasets based on absolute lengths of interaction times, i.e., the first 1000ms, 2000s, 3000ms, etc. of each trial. These datasets can therefore be seen as more realistic in terms of evaluating classification accuracies *while* a user is interacting with (i.e., looking at) the visualization. One particular challenge in using this approach is that users do not necessarily complete tasks in the same time. Thus, as we train classifiers with increasing absolute time intervals, there is the question of how to deal with participants that have already finished a task prior to the current time interval, specifically whether they should be included in the training sets for time intervals longer than their completion time or not. Based on comparative tests (which we will not present in the paper), we found that we can obtain better results by retaining the full data set across all classification time slices, even if some users in the training set have finished before a specific time cut-off. One simple explanation for this result could be that the reduction of the training set may leave the classification algorithm with too few examples to learn from, particularly for longer trials.

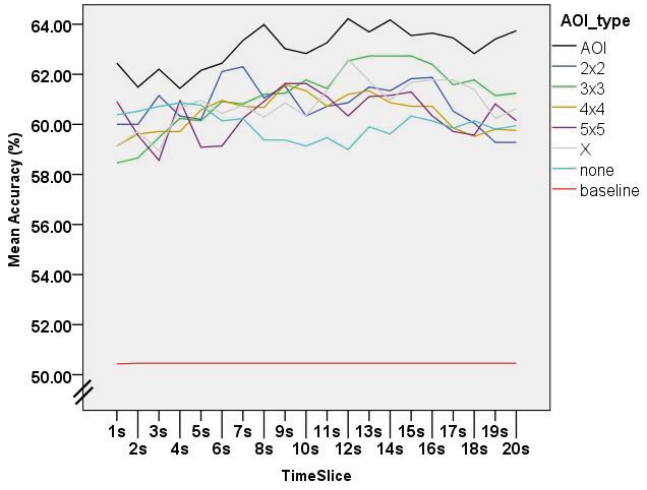

Figure 7. Classification accuracy using absolute time to generate eye tracking features

Similar to the percentage-based experiments, we computed the classification accuracy over incremental time slices (in this case 20 absolute-time intervals, from 1000ms to 20000ms) for each of the seven AOI sets. The resulting overtime trends are shown in Figure 7. In general, we see that the *Visualization-Specific* classifier generates the best accuracy right from the beginning, and that all classifiers comfortably outperform the baseline classifier, reaching accuracies of up to 64%. Also, some classifiers already reach over 60% accuracy after only seeing a few seconds of eye tracking data, hence providing very encouraging results for our long-term goal of developing user-adaptive systems during a user's system usage. Similarly, while the *Visualization-Specific* classifier performs best overall, it is worth noting that all classifiers perform in the region of 60% across all time intervals (including the classifier with no AOI information), which could potentially be combined with other sources such as interaction or mouse-tracking data to achieve even higher accuracies (and hence drive a user-adaptive system).

We also compared the average over time performance between both the *percentage-based* and the *absolute-time-based* segmentation approaches. We ran a 2 (percent/absolute) by 8 (AOI-Type) ANOVA with classification accuracy as the dependent measure. Results indicated no significant differences between the percentage and absolute time classifiers ($F_{1,44}$= 1.04, p = .314). The main effect of AOI-Type, however, is again significant ($F_{7,2357}$= 408.8, p < .001), and Bonferroni adjusted pairwise comparisons yield identical results as reported in the previous subsection, i.e. showing that *Visualization-specific* classifiers perform best, and that the 3x3 is the best performing *Generic* AOI set. The lack of a statistically significant difference between percentage and absolute time slices is also interesting, because it indicates that the percentage-based approach (often used in prior research) is not gaining an unrealistic/unfair advantage from the fact that it potentially encodes the task length in the eye gaze features (e.g. longer trials having a higher number of fixations in the first 10%, hence being indicative of a *during* trial).

Feature Selection

In addition to examining the different accuracies of classifiers, we are also interested in studying which features are most predictive for skill acquisition classification (research question 2). In particular, we want to qualify the differences in behavior *during* and *after* skill acquisition. We therefore ran feature selections for each time slice using the correlation-based feature selection (CFS) method described in [12], which picks the top attributes (on average 10 for our data sets) based on the individual predictive ability of each feature along with the degree of redundancy between them. We then generated frequency tables across all time slices to analyze which features were chosen most often. We repeated this process for both percentage-based and absolute-time classifier sets. Across each of the classifiers, there were five eye-gaze features that consistently appeared across time slices, and which were particularly influential during the early stages. In particular, high values for *standard deviation of relative path angles*, *sum of absolute path angles*, as well as *sum of relative path angles* were found to be indicative of users *during* skill acquisition. This supports prior hypotheses that state that 'relative angles within a scan path indicate the directness of scanning, and therefore the complexity or uncertainty of the task and page layout' [10], since untrained users would

typically be more uncertain about how to read the visualization and hence perceive the task to be more complex (a result that was also found in [4]). Similarly, we found that high values of *sum of path distances* and *mean of absolute path angles* were found to be predictive of trials *after* skill acquisition. This may indicate that trained users have more confident gazes, since their individual paths are longer and hence more 'assertive'. During longer time slices (especially after 15000 ms), *Visualization-Specific* AOIs became increasingly important for classification, for example with *during* trials spending significantly more time in the *question text*.

Overall, these results explain why even the classifier with no AOI information (i.e. '*none'*) performed comparably well for early time slices (since the five features above were all AOI independent), whereas the *Visualization-Specific* classifier became increasingly accurate during the later stages of trials.

DISCUSSION & FUTURE WORK

This paper has presented our initial steps towards facilitating adaptive help to ease a user's learning curve.

First of all, we have verified that even with simple visualizations, it is possible to identify a significant difference in performance based on the amount of practice a user has had with the system, regardless of other experimental conditions. We have tracked this performance difference using a learning curve, which allowed us to split user trials into two general stages of skill acquisition (*during* and *after* skill acquisition).

Second, this paper has explored the feasibility of building an online classifier that aims to make predictions *while* a user is using a visualization for a short task. We have shown that from the outset, our gaze-based classifiers outperform a simple baseline, and that even after observing only a few seconds of gaze data, we can make predictions with up to 60% accuracy. While these accuracies may not be high enough yet for driving a user-adaptive system, it is worth noting again that in this paper we have solely used gaze data, and that the combination of this data with complementary input features such as interaction data, mouse-tracking data, or other user characteristics is likely to improve prediction accuracies (as shown in [15]).

Third, in order to investigate the generalizability of detecting user skill acquisition across different types of visualizations and other interfaces, we have provided initial results on the relative performance of a *Visualization-Specific* AOI set (less generalizable) compared to *Generic* AOIs (applicable to any user interface). Results of this analysis have shown that, while visualization-specific AOI sets are the most predictive overall (as to be expected), reasonably accurate predictions can be achieved without detailed knowledge of the visualization/interface. In particular, we showed that the generic 3x3 AOI grid performed best overall (compared to more/less fine-grained grids), indicating that there is a trade-off in terms of having too few AOIs versus too many AOIs.

While these are already encouraging results, there are many alternative methods that we plan to apply in future work in order to build realistic user-adaptive systems. For example, while our time-slices are currently of a 'cumulative' nature, we can investigate 'sliding window' classifiers to see if there are particular segments (e.g., in the middle of a user's system usage) that are more discriminative. Similarly, in this paper we have only attempted to predict a user's skill level based on individual trials (with trials consisting of very short periods of 'interaction'), which is arguably a very difficult task for data-driven classification. An alternative that we envision is to classify based on longer periods of user gaze data, for example through combining a number of trials to form longer interaction periods. Another alternative could be to track gaze behavior over such longer periods, and then classify users based on particular *changes* in behaviors.

Similarly, while the simple two-way split has given us some initial insights into the feasibility of user classification, as well as the behavioral differences that accompany a user's skill level, there are a number of potential enhancements to explore. For example, since the learning curve follows a typical power law [18] (see Figure 3), we may investigate alternative splits (e.g., labeling the first 20 trials as *during* and the rest as *after*) to tease out even stronger and more discriminatory features. Likewise, we expect the effect of the learning curve to be stronger when we move to more complex or unfamiliar types of visualizations or tasks. This will also allow us to investigate how different learning curves map to differences in behavior.

In terms of generalizability, we intend to reuse the *Generic* AOIs in future experiments, where we will combine data from different visualizations and interfaces to investigate general trends of user behavior. If classifiers based on such heterogeneous data sets are indeed able to predict user skill acquisition, our methods would hence point towards a general method to identify user skill across visualization systems in general.

Lastly, future research challenges also lie in devising adaptive mechanisms that can actually aid users during the skill acquisition phase. Such challenges include, for example, choosing the type of adaptation to use (e.g., tooltips vs. reducing interface functionalities), as well as analyzing the relative benefits and drawbacks of such adaptations (e.g., adaptation effectiveness vs. intrusiveness).

REFERENCES

1. Adams, J.A. Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin 101*, 1 (1987), 41–74.

2. Amar, R., Eagan, J., and Stasko, J. Low-Level Components of Analytic Activity in Information Visualization. *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, IEEE Computer Society (2005), 15–.
3. Bednarik, R., Eivazi, S., and Vrzakova, H. A Computational Approach for Prediction of Problem-Solving Behavior Using Support Vector Machines and Eye-Tracking Data. In Y.I. Nakano, C. Conati and T. Bader, eds., *Eye Gaze in Intelligent User Interfaces*. Springer London, London, 2013, 111–134.
4. Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J.M., Azevedo, R., and Bouchet, F. Inferring Learning from Gaze Data during Interaction with an Environment to Support Self-Regulated Learning. In H.C. Lane, K. Yacef, J. Mostow and P. Pavlik, eds., *Artificial Intelligence in Education*. Springer Berlin Heidelberg, 2013, 229–238.
5. Bunt, A., Conati, C., and McGrenere, J. Supporting interface customization using a mixed-initiative approach. ACM Press (2007), 92.
6. Carenini, G., Conati, C., Hoque, E., Steichen, B., Toker, D., and Enns, J.T. Highlighting Interventions and User Differences: Informing Adaptive Information Visualization Support. (2013), (accepted).
7. Conati, C. and Merten, C. Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. *Knowledge-Based Systems 20*, 6 (2007), 557–574.
8. D'Mello, S., Olney, A., Williams, C., and Hays, P. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies 70*, 5 (2012), 377–398.
9. Dillon, A. Spatial-semantics: How users derive shape from information space. *Journal of the American Society for Information Science 51*, 6 (2000), 521–528.
10. Goldberg, J.H. and Helfman, J.I. Comparing information graphics: a critical look at eye tracking. *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization*, ACM (2010), 71–78.
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl. 11*, 1 (2009), 10–18.
12. Hall, M. Correlation-based Feature Selection for Machine Learning. 1999. http://www.cs.waikato.ac.nz/~mhall/thesis.pdf.
13. Jameson, A. The human-computer interaction handbook. In J.A. Jacko and A. Sears, eds., L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003, 305–330.
14. Kardan, S. and Conati, C. Exploring gaze data for determining user learning with an interactive simulation. *In: Proc. of UMAP, 20th Int. Conf. on User Modeling, Adaptation, and Personalization*, (2012), 126–138.
15. Kardan, S. and Conati, C. Comparing and Combining Eye Gaze and Interface Actions for Determining User Learning with an Interactive Simulation. *In: Proc. of UMAP, 21st Int. Conf. on User Modeling, Adaptation and Personalization*, (2013).
16. Lewandowsky, S. and Spence, I. Discriminating Strata in Scatterplots. *Journal of the American Statistical Association 84*, 407 (1989), 682–688.
17. Linton, F. and Schaefer, H.-P. Recommender Systems for Learning: Building User and Expert Models through Long-Term Observation of Application Use. *User Modeling and User-Adapted Interaction 10*, 2-3 (2000), 181–208.
18. Logan, G.D. Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition 18*, 5 (1992), 883–914.
19. McDonald, S. and Stevenson, R.J. Navigation in hyperspace: An evaluation of the effects of navigational tools and subject matter expertise on browsing and information retrieval in hypertext. *Interacting with Computers 10*, 2 (1998), 129–142.
20. Pascual-Cid, V., Vigentini, L., and Quixal, M. Visualising Virtual Learning Environments: Case Studies of the Website Exploration Tool. IEEE (2010), 149–155.
21. Qu, L. and Johnson, W.L. Detecting the Learner's Motivational States in An Interactive Learning Environment. *Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, IOS Press (2005), 547–554.
22. Saraiya, P., North, C., and Duca, K. An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. *IEEE Transactions on Visualization and Computer Graphics 11*, 4 (2005), 443–456.
23. Speelman, C. and Kirsner, K. *Beyond the Learning Curve*. Oxford University Press, 2005.
24. Steichen, B., Carenini, G., and Conati, C. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM (2013), 317–328.
25. Toker, D., Conati, C., Carenini, G., and Haraty, M. Towards adaptive information visualization: on the influence of user characteristics. *Proceedings of the 20th international conference on User Modeling, Adaptation, and Personalization*, Springer-Verlag (2012), 274–285.
26. Toker, D., Conati, C., Steichen, B., and Carenini, G. Individual user characteristics and information visualization: connecting the dots through eye tracking. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 295–304.
27. Velez, M.C., Silver, D., and Tremaine, M. Understanding visualization through spatial ability

differences. *IEEE Visualization, 2005. VIS 05*, (2005), 511–518.

28. Zhu, Y. Measuring effective data visualization. In *Advances in Visual Computing*. Springer, 2007, 652–661.
29. Ziemkiewicz, C., Crouser, R.J., Yauilla, A.R., Su, S.L., Ribarsky, W., and Chang, R. How locus of control influences compatibility with visualization style. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, (2011), 81–90.