Toward User-Adaptive Visualizations: Further Results on Real-Time Prediction of User Cognitive Abilities from Action and Eye-Tracking Data

by

Md Abed Rahman

B.Sc., Islamic University of Technology, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

June 2017

Abstract

Previous work has shown that some user cognitive abilities relevant for processing information visualizations can be predicted from eye-tracking data. Performing this type of user modeling is important for devising user-adaptive visualizations that can adapt to a user's abilities as needed during the interaction. In this thesis, we contribute to previous work by extending the type of visualizations considered and the set of cognitive abilities that can be predicted from gaze data, thus providing evidence on the generality of these findings. We also evaluate how quality of gaze data impacts prediction. Finally, we further extend previous work by investigating interaction data as an alternative source to predict our target user characteristics. We present a formal comparison of predictors based solely on gaze data, on interaction data, or on a combination of the two.

Lay Summary

This thesis aims in finding the feasibility of modeling users in a commercial system, one that would facilitate the development of systems that adapt to users' needs. This thesis explores three potential sources of information for user modeling: eye-tracking, interaction and a combination of both. First contribution of this thesis is to provide a first inspection of the tradeoff between having sufficient amount of eye-tracking data for user modeling and having good quality data, since, eye-tracking data is inherently noisy. Another contribution of this thesis is to explore the use of user interaction and a combination of eye-tracking and user interaction data for user modeling. A comparative analysis performed on the feasibility of using the aforementioned three source of information is also presented in this thesis. Finally, this thesis also provides insights on the feasibility of real-time user modeling, a pre-requisite for the development of adaptive systems.

Preface

The study presented in Chapter 3 was not designed by the author of this Thesis. This includes decisions such as: which user characteristics were selected, design of the task, which tasks to use, the generation of raw data. Generating the subsequent datasets from raw data used for this thesis has been done by the author of this Thesis.

A version of the research reported in Chapters 4 and 5 has been accepted to be published as: Conati, C., Lallé, S., Rahman, M. A., Toker, D. (2017) Further Results on Predicting Cognitive Abilities for Adaptive Visualizations. To appear in IJCAI 2017

Table of Contents

Abstract ..... **ii**

Lay Summary ..... **iii**

Preface ..... **iv**

Table of Contents ..... **v**

List of Tables ..... **vii**

List of Figures ..... **ix**

Acknowledgements ..... **x**

Dedication ..... **xi**

Chapter 1: Introduction **1**

Chapter 2: Related Work **5**

Chapter 3: Dataset **9**

3.1 MetroQuest (MQ) 9

3.2 User Study 10

Chapter 4: Eye-Tracking Data Processing **13**

4.1 Data Windows 13

4.2 Eye-Tracking Features 14

4.2.1 Gaze Features 14

4.2.2 Pupil features 14

4.2.3 Head Distance features 15

4.3 Eye-tracking Data Validity Thresholds 16

Chapter 5: Classification Experiments and Results **18**

5.1 Feasibility of Predicting User Characteristics 18

5.2 Effects of Data Validity Threshold 19

5.3 Further Results for VisWM 22

Chapter 6: Prediction using Action Features 24

6.1 Low-level action features 24

6.2 High-level action features 25

6.3 Feasibility of Predicting User Characteristics using action features 28

Chapter 7: Comparing the data sources 32

7.1.1 Finding the Best Data Source and Feasibility of Early Predictions using Best Data Source 33

Chapter 8: Conclusion 36

Bibliography 39

Appendices 42

Appendix A - Correlation Analysis 42

A.1 High Level Proportion Features 42

A.2 Low and High-Level Features 43

List of Tables

Table 4.1 Summary statistics on user-characteristics scores obtained……. 24

Table 4.2 Set of features considered for classification……. 26

Table 5.1 Accuracies (averaged across data windows) and F-statistics for the main effects of classification algorithm (RF, LF and baseline) for each of PS, SpM, VisScan and VisWM, and for each data validity threshold tested. † indicates that RF or LB significantly beat the corresponding baseline. Bold indicates that RF significantly beat LB ……..... 30

Table 5.2 Comparison of RF accuracy with different data validity thresholds, † indicates that RF significantly beat the corresponding baseline 31

Table 5.3 Windows for which the highest accuracies are achieved for the given validity thresholds…… 32

Table 6.1 Set of low-level Action features considered for classification……... 35

Table 6.2 Set of high-level actions……. 36

Table 6.3 Set of high-level Action features considered for classification……... 38

Table 6.4 Comparison of classification accuracies for the User characteristics that can be predicted better than baseline, † indicates that RF significantly beat the corresponding baseline. …… 41

Table 6.5 Windows for which the highest accuracies are achieved for the chosen classifier (RF) …… 42

Table 7.1 Windows for which the highest accuracies are achieved for the chosen classifier (RF)…… 45

Table 7.2 Windows for which the highest accuracies are achieved for the chosen data source.. 46

Table A.1.1 Pearson’s correlation coefficient values for proportion of time and proportion of actions features ……. 54

Table A.1.2 Pearson’s correlation coefficient values for high level frequency and low-level features ……. 56

List of Figures

Figure 3.1 MQ interface used in the study …… 22

Figure 4.1 Number of participants with valid eye-tracking data as the strictness of validity threshold is increased from 0.5 to 1. 28

Figure 5.1 Accuracy of RF classifiers using data with three different validity thresholds for: PS, SpM, VisScan and VisWM..... 34

Figure 6.1 Average number of each type of high-level actions performed per user (blue bars) and number of different users who performed each action type at least once (red bars) 38

Figure 6.2 Results for PS, SpM, VisWM and VisScan with action features…… 41

Figure 7.1 Comparison of accuracies for action, eye-tracking and mixed featuresets for PS, SpM, VisWM and VisScan…… 45

Acknowledgements

First and foremost, I would like to thank my supervisor, Cristina Conati, for accepting me as a master's student. She has always been direct in her feedback and efficient in her approach towards research. Her guidance helped me get the proper exposure to academic research that I wished to have as a master's student. She also managed to find some time to listen to my concerns when I encountered difficulties. Thank you.

Next, I would like to thank Dereck Toker for being a friend as well as a dependable colleague; someone I could always count for his wisdom and professional advice.

I would also like to thank Sebastién Lallé, who helped me immensely to get started on my journey to working on my master's thesis by making the transition from where his previous work on the project finished and where my contribution in the project started, as smooth as possible. He is truly an inspiring colleague with great work ethic.

Also, I would like to thank Giuseppe Carenini for his feedback and insights provided during group meetings and inspiring comments throughout the course of my master's research work.

Again, thanks to Jim Little for agreeing to be the second reader of this thesis.

Last but not least, thanks to my family and a special someone for their unconditional love and support. Thank you.

For my dad, my best friend ever.

Chapter 1: Introduction

Information visualization (InfoVis) is a thriving area of research that takes advantage of the strength of human perception to facilitate the analysis of complex data. There is mounting evidence that several cognitive abilities and traits can influence users' visualization experience both in terms of overall task performance (e.g., [Toker et al. 2012; Ziemkiewicz et al. 2011]) as well as in how well users can process specific elements of a visualization (e.g., [Jang et al. 2014; Ooms et al. 2014; Toker et al. 2013]). These findings support the value of having user-adaptive visualizations, i.e., intelligent interfaces that learn about their users (user modeling) and adapt the visualization to meet each user's needs in real time [Conati et al. 2015].

Previous work has shown that some of the user cognitive abilities known to be relevant for processing information visualizations – perceptual speed, visual working memory (WM), and verbal WM – can be predicted in real time from eye-tracking data [Gingerich and Conati 2015; Steichen et al. 2014]. These findings provided encouraging evidence on the feasibility of performing user modeling; necessary for user-adaptive visualization. However, this previous work focused on users processing either bar graph or radar graph visualizations, to perform fictional question-answering tasks abstracted from any real usage context. We contribute to this previous work by replicating results on the real-time prediction of perceptual speed and visual WM, for users working with two very different types of visualizations (a deviation chart and a map-based visualization), embedded in a commercial application designed to engage the public in decision making related to urban planning. We also show that real-time prediction is feasible for two other cognitive abilities (visual scanning and spatial memory), not previously considered and relevant for processing the new visualizations we investigated. These results are an

important step for advancing research on user-adaptive visualizations from initial proof of concepts to more generalizable findings.

We also evaluate how quality of eye-tracking data impacts prediction accuracy. There are promising results on the value of eye-tracking data for predicting a variety of user states and abilities in user modeling (e.g., [Bednarik et al. 2013; Gingerich and Conati 2015; Jaques et al. 2014; Kardan and Conati 2012; Ooms et al. 2014; Lallé et al. 2016b]). However, eye-tracking data can be rather noisy, due to several factors such as user eye physiology (e.g., wearing glasses), excessive movement, design of the eye tracker, etc. [Holmqvist 2011]. Existing user-modeling research has mostly dealt with this problem either by adopting usually laborious procedures to increase data quality during data collection, or by discarding too-noisy data. In addition to being time consuming, these approaches provide results that have limited generalizability to real-world settings, where eye-tracking data is bound to be noisy. In this thesis, we show that relatively noisy eye-tracking data can still be used for prediction, thus providing encouraging, albeit preliminary, evidence on the applicability of our findings to real-world scenarios.

Additionally, we show that action data i.e., data about the user's interaction with the system can also be used to predict said user characteristics. There have been positive results about using action data for predicting user's cognitive states during interaction with educational software [Kardan and Conati 2013], framework designed to capture fatigue [Pimenta et al. 2013] and InfoVis [Lallé et al. 2016a] . Kardan and Conati [2013] used action data to predict user learning with an interactive simulation for AI algorithms. Pimenta et al. [2013] showed that it is possible to predict mental fatigue using keyboard and mouse interaction data that were logged using a simple application that ran in the background, while the user performs day to day tasks in a

computer. Lallé et al. [2016a] showed that it is possible to predict user confusion using interaction data during user's interaction with a visualization that supports decision making. In this thesis, we explore the value of action data to predict our target set of user characteristics during interaction with MQ because eye-tracking data might not always be available, especially in real-world applications. Action data, on the other hand, is always available and easier to track. Finally, we investigate whether fusing eye-tracking and action data can provide better prediction accuracy than each individual data source in isolation. Namely, we derive a third combined set of features, and perform an analysis for comparing the predictive capabilities of eye-tracking, action and combined data sources. Exploring the value of combining eye-tracking and action data has been done by Kardan and Conati [2013] and Lallé et al. [2013] in the domains described earlier, to capture learning and confusion, respectively. Both Kardan and Conati [2013] and Lallé et al. [2013] show that the fused data source is superior in predictive capabilities than the individual sources [i.e., action and eye-tracking]. In contrast, our results show that for 3 [out of 4] user characteristics, eye-tracking performs as well as or better than the fused feature set; a result which is different than before. This finding indicates that the value of combining eye and action data depends on the user states and traits to be predicted, and possibly on the interaction task, calling for future research to better understand how prediction target and interaction task affect data source value. Additionally, our analysis indicates that, while for some cognitive abilities eye-tracking and the fused feature set give the best predictive accuracy, action features do best at early predictions, namely they have better accuracy than the other two data sources in earlier stages of interaction. Thus, our work provides evidence on the possible merits of leveraging these data sources separately instead of always fusing them to get better results. Specifically, given our long-term goals of using these predictions for providing real-time adaptive support for

visualization processing, these results call for further investigation of the tradeoff between giving early predictions and accuracy of the interventions.

The rest of the thesis starts with an overview of related work, followed by a description of the study that generated the dataset that we used for this research. Next, we explain the processing of the eye-tracking data obtained from the study, classification experiments performed on eye-tracking data and analysis of their results. Then, we repeat the process for action data by explaining the processing of action data, describing the classification experiments done and presenting the analysis of the results obtained for action data. Finally, we describe the generation of the fused data source followed by a comparative analysis of predictive capabilities between eye-tracking, action and combined data sources.

Chapter 2: Related Work

There is increasing interest in integrating AI and InfoVis research to devise user-adaptive visualizations that can support the specific needs on each individual user. Such user-adaptive visualizations would be especially beneficial as complex visualizations are becoming increasingly used by broad audiences, not only in professional settings, but also for personal usage (e.g., for monitoring health and fitness, interactions in social media, and home resources consumption) [Huang et al. 2015] .

Previous work suggests that there is value in user modeling for the development of user-adaptive visualizations [Domik and Gutkauf 1994] . Work to date has leveraged various sources of information to perform user modeling that would facilitate the development of user adaptive interfaces. One such source is eye-tracking data. Eye-tracking has been shown to be a good predictor of other emotional or attentional states such as mind wandering while reading [Bixler and D'Mello 2015] as well as boredom, curiosity and excitement, while learning with educational software [Jaques et al. 2014; Muldner et al. 2010] . There has also been research on leveraging eye-tracking data for predicting user cognitive abilities (namely perceptual speed, visual working memory (WM), and verbal WM) that have been shown to be relevant for processing information visualizations based on bar and radar graphs. Other work has shown that these cognitive abilities impact the processing of specific elements of bar and radar graphs. For instance, lower levels of perceptual speed can result in slower processing of legend and labels in bar graphs [Toker et al. 2013], suggesting that these users might benefit from personalized interventions geared toward facilitating legends and labels processing if the visualization can detect in real-time that they have low perceptual speed.

[Lallé et al. 2017] showed similar impact of cognitive abilities on how users process two different types of visualizations in MetroQuest (MQ), a commercial system designed to support decision making for environmental problems. These results suggest that adaptive interventions based on user cognitive abilities could enhance user experience with MQ. One of the goals of this thesis is to investigate whether previous results on predicting cognitive abilities in real-time with eye-tracking data can be reproduced on data collected with MQ, both for some abilities already seen in previous work and for new ones specific to [Lallé et al. 2017].

Work in user modeling has typically handled noise in eye-tracking data by discarding users or trials with too noisy data. An alternative approach adopted in [Kardan and Conati 2012] involved monitoring data generated by each participant in real-time and requesting adjustments when noisy data was observed, e.g., asking the participants to move less. There are techniques in eye-tracking research designed to reduce noise without discarding data, such as removing artifacts responsible for noise, e.g., blinks [Holmqvist 2011], or smoothing noisy gaze datapoints based on the latest clean ones [Špakov 2012]. However, there is limited understanding on how well these techniques scale up with the increase of noise in the data. Also, little is known about the need to use these techniques, i.e., the need for cleaner data for good predictions. Thus, a second goal of this thesis is to provide insights on how noise in eye-tracking data affects the prediction in a user modeling task.

Another stream of research in user-adaptive visualizations has leveraged user interaction data for user modeling [Ahn and Brusilovsky 2013; Gotz and Wen 2009; Lim et al. 2015; Liu et al. 2010; Mouine and Lapalme 2012; Nazemi et al. 2014; Yelizarov and Gamayunov 2014]. Ahn and Brusilovsky [2013] worked on proposing and implementing an adaptive visualization based search system using actions, one which provided tailored search results based on modeling users

based on their interests. Gotz & Wen [2009] contributed to the development of a behavior-driven visualization recommendation system. Their idea was to monitor a user's interaction data in order to define and detect suboptimal usage patterns, and then adapt to these patterns by recommending alternative visualizations to the user. Lim et al. [2015] measured cognitive stress in students by using mouse and keyboard dynamics. They have shown that mouse click rate, mouse speed etc. change based on user's cognitive stress levels. Liu et al. [2010] used mouse click logs to perform a detailed analysis of user browsing behaviors. Mouine and Lapalme [2012] worked on providing personalization of visualization using interface actions; this was achieved by automatic user profiling. Nazemi et al. [2014] used actions to develop an approach to find the average user of a visual environment to provide adaptation. They also used this user model to measure the distance between an average user and individual users to detect anomalies in user behaviors. Yelizarov and Gamayunov [2014] tracked mouse and keyboard events to predict users' level of cognitive load and adjusted the amount of information to be displayed accordingly.

Another goal of this thesis is to compare gaze, action and combination of gaze and action as predictors of our target user characteristics [namely perceptual speed, visual working memory (WM), Visual scanning, verbal WM and Visual Literacy; [see Section 3.2 for details]. Previous work has done similar comparisons for predicting user cognitive states [Kardan and Conati 2013; Lallé et al. 2016a]. Kardan and Conati [2013], used both eye-gaze and action data to modeling user's learning in Interactive Simulations(IS). They trained classifiers on eye-tracking only, action only and on fused data [action and eye-tracking]. Finally, they also did a comparative analysis of the predictive capabilities of gaze data, interface actions and a combination of these two sources of user interaction data. Lalle et. al [2016a] worked on predicting confusion that

occur during working with an interactive visualization named ValueCharts. They showed that a combination of gaze and action features give the best predictions for confusion. In this thesis, we do a similar analysis for eye-tracking, action and fused [action and eye-tracking] for the user characteristics reported earlier.

Chapter 3: Dataset

The data used in this thesis was collected during a user study (mentioned in the related work and fully described in [Lallé et al. 2017]) that investigated the impact of individual differences on user experience and gaze behavior with MetroQuest (MQ). Here we provide a brief summary of MQ and of the study, sufficient for the purposes of this thesis.

3.1 MetroQuest (MQ)

MQ supports rapid customization of a set of standardized screens that guide users through the process of learning about a target decision problem, defining their preferences over the decision factors, exploring various outcome scenarios and generating their decision.

The MQ interface used in this study (Figure 3.1) addresses the problem of building a new transportation system to our campus. This is a real project currently studied by the City, and it has generated substantial controversy on which of the proposed transit scenarios (light rail, rapid rail, or a combination of both) should be selected. MQ allows users to rank their priorities for seven factors that are affected by the transit decision (e.g., travel time saving to campus, wait time, frequency of stops, reduction in auto trips and pollution). Then MQ shows how each of the proposed transit scenarios affects the decision factors by displaying two complementary visualizations, a deviation chart and a map.

The deviation chart (Figure 3.1, upper right hand side of the screen) indicates whether the value of each factor improves (green arrow) or worsens (red arrow) compared to the current situation. Arrow size shows the magnitude of the difference. The map (Figure 3.1, bottom) displays factual information on the planned transit scenario (e.g., the route, stop locations) as well as the actual values for factors by means of map keys (e.g., time savings are reported at stops along the route).

A button allows opening the legend for the keys of the map. Users can view and compare the different scenarios by using the tabs shown at the top left of Figure 3.1. A short textual description of each scenario is provided below the tabs. Users can rate a scenario by using the scale shown below the textual description.

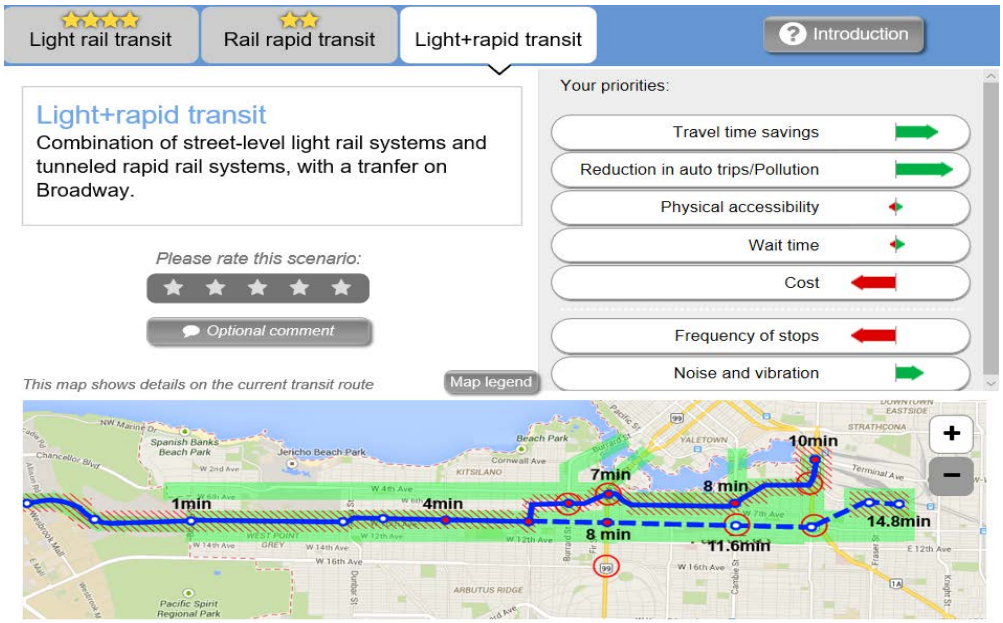

Figure 3.1 MQ interface used in the study.

3.2 User Study

In the study, 166 participants were invited to use MQ to learn about and provide their preferences on the transit scenarios described in the previous subsection. They used a mouse to click on various parts of the screen. They also used keyboards to provide feedback about each individual transit scenario. This mimics how MQ is often used in public settings, such as in information kiosks, where target communities engage in one-time interaction with MQ. Participants were recruited among the population living or working on campus, and thus undoubtedly have a real interest in the task. During the interaction with MQ, participants' gaze

was tracked with the Tobii T120, a non-intrusive camera-based eye tracker embedded in the study monitor. Prior to engaging in the study task, each participant underwent a standard calibration phase with the eye tracker. The Tobii T120 also tracks pupil size, which we include in the dataset for predicting user characteristics (see Section 4.2). To compensate for physiological differences in pupil size among users, pupil diameter baselines were collected for each participant by having them stare at a blank screen for ten seconds. To avoid possible confounds of pupil size due to lighting changes, the study was administered in a windowless room with uniform lighting. After completion of the task with MQ (mean completion time = 4min 45sec, st. dev. = 2min 01sec), participants filled a postquestionnaire about their experience with MQ. Lastly, each participant took a battery of tests to measure 12 user characteristics. In this thesis, we focus on five of them:

- Perceptual speed (PS), a measure of speed when performing simple perceptual comparisons [Ekstrom et al. 1976];
- Visual working memory (VisWM) a measure of storage and manipulation capacity of shapes and colors of visual objects [Fukuda and Vogel 2009];
- Spatial memory (SpM), a measure of storage and manipulation of the spatial arrangement of objects [Ekstrom et al. 1976];
- Visual scanning (VisScan, a measure of the capacity to actively find relevant information in our surroundings quickly and efficiently [Ekstrom et al. 1976].
- Visualization literacy (VisLit, the "ability to use well-established data visualizations to handle information in an effective and efficient manner" [Boy et al. 2014][1].

[1] *PS*, *SpM*, *VisScan* and *VisWM* were collected using the Kit of Factor-Referenced Cognitive Tests [Ekstrom et al. 1976]. The Fukuda & Vogel's test [Fukuda and Vogel 2009] was used for *visual WM*. *VisLit* was collected using a formal test that has been recently proposed [Boy et al. 2014].

We focus on these five characteristics because a previous analysis [Lallé et al. 2017] on this dataset has shown that they are the ones influencing user experience and gaze behavior with the MQ visualizations. Specifically, VisWM affects user preference between chart and maps, i.e., users with high VisWM preferred the deviation charts over the maps. SpM influenced perceived visualization usefulness, i.e., users with lower SpM found the deviation chart less useful than users with higher SpM. SpM, along with PS, VisScan, and VisLit, also influenced gaze behaviors related to making comparisons between visualizations across scenarios, i.e., users with lower levels of these characteristics made fewer visual comparisons than users with higher levels. These results indicate that it could be beneficial to predict in real time whether MQ users have lower or higher levels of the aforementioned abilities and provide adaptive support accordingly. Such support could include for instance, interventions designed to make deviation charts more useful for users predicted to have low SpM, or to facilitate comparisons between scenarios for users with low levels for the relevant abilities. In the next sections, we discuss the eye-tracking data and machine learning experiments we used to ascertain if making such predictions is possible.

Chapter 4: Eye-Tracking Data Processing

We leverage the eye-tracking data collected during the user study described in the previous section to build classifiers that can predict binary labels of the five user characteristics reported in Section 3.2. The binary labels were generated by dividing participants into "High" and "Low" groups for each characteristic (e.g., High and Low perceptual speed), based on a median split on the test scores from the study. Table 4.1 shows a table that includes summary statistics on the test scores for various user characteristics; obtained from the previously reported user study.

User Characteristic	Max	Min	Mean	Standard Deviation
PS	63	13	39.6	9.1
SpM	22	0	11.2	5.3
VisWM	5.6	-1.2	2.1	1.2
VisLit	1	-1.67	.41	.5
VisScan	40	0	24.6	8.1

Table 4.1 Summary statistics on user-characteristics scores obtained

4.1 Data Windows

To simulate the real-time prediction of user characteristics, we generated ten data windows corresponding to incremental percentages (10%, 20%... up to 100%) of eye-tracking data during interaction with MQ. This approach allows us to verify how early during the interaction with MQ the target user characteristics can be predicted. Investigating prediction timing is of prime importance for our goal of providing adaptive support in real-time to users with specific characteristics. Early adaptation is especially important in systems like MQ that typically target one-time users for a short period of time, and thus need to be quickly understandable. To predict

user characteristics, we generated a battery of eye-tracking features (described next) at each data windows.

4.2 Eye-Tracking Features

The Tobii eye tracker captures user's gaze samples, i.e., where the user is looking on the screen, at 120 Hz. Then fixations (gaze maintained at one point on the screen) and saccades (quick eye movement between two fixations) are derived from gaze samples. For every recorded gaze sample, the eye tracker also captures pupil size and distance from user's head to the screen. From all these measures (Gaze, Pupil, and Head Distance) we derived a set of features listed in Table 4.2 that we leveraged to predict user characteristics during interaction with MQ. We used EMDAT (https://github.com/ATUAV/EMDAT), an eye-tracking data analysis toolkit, to generate these features.

4.2.1 Gaze Features

EMDAT generated the gaze features listed in Table 4.2 (part a) by calculating various summary statistics (e.g., sum, mean) over a user's fixations and saccades. These statistics are computed for gaze movements over the whole interface, generating the gaze features labelled as Overall Gaze Features in Table 4.2 (part a), or they can be computed over specific areas of interest (AOI) in the MQ interface, generating the AOI Gaze Features in Table 4.2 (part a). There are four AOIs defined over four regions of MQ (which is shown Figure 3.1): Description of the transit scenario; Deviation chart; Map; Legend of the map.

4.2.2 Pupil features

Pupil sizes were adjusted using the pupil baseline collected during the study, following [Iqbal et al. 2005]. Using EMDAT, we computed a set of summary statistics on user-adjusted pupil size, suitable for describing fluctuations of this measure over the course of the interaction with MQ.

These include min, max, mean, and std. dev. of users' pupil sizes in each data window [see Table 4.2 (part b)]. We also included the measure of a user's pupil at the beginning and the end of the current data window (Pupil width at the first and last fixation in the data window Table 4.2 (part b), as a way to capture pupil size variations between the start and the end of each window.

4.2.3 Head Distance features

Head distance is obtained by averaging the distances from both eyes to the screen. We used EMDAT to compute the same set of statistics as for pupil size [see Table 4.2 (part c)], as described above.

a) Gaze Features (68)
Overall Gaze Features (12):
Fixation rate
Mean & Std. deviation of fixation durations
Mean & Std. deviation of saccade length
Mean, Rate & Std. deviation of relative saccade angles
Mean, Rate & Std. deviation of absolute saccade angles
Mean saccade velocity
AOI Gaze Features for each AOI (56):
Fixation rate in AOI
Longest fixation in AOI, Time to first & last fixation in AOI
Proportion of time, Proportion of fixations in AOI
Number & Prop. of transitions from this AOI to every AOI
b) Pupil Features (6) and c) Head Distance Features (6)
Mean, Std. deviation, Max., Min. of pupil width/head distance
Pupil width/head distance at the first and last fixation in the data window

Table 4.2 Set of features considered for classification.

4.3 Eye-tracking Data Validity Thresholds

As described in the introduction, we want to investigate if and how quality of eye-tracking data influences the real-time prediction of our target user characteristics. The Tobii eye tracker marks each gaze sample as valid or not. Too many invalid gaze samples may make the data for a given user unreliable to represent their gaze, pupil and head behaviors. However, there are no established guidelines to ascertain how many invalid samples are too many. One could be conservative and include only users with small percentages of invalid samples, but this can severely reduce the size of the dataset. To illustrate, Figure 4.1 shows the percentage of study participants in our dataset with a proportion of valid gaze samples higher than validity thresholds ranging from 0.5 to 1. Setting a validity threshold of 0.9 (i.e., including participants with at least 90% of valid gaze samples) would exclude about 40% of users in our dataset.

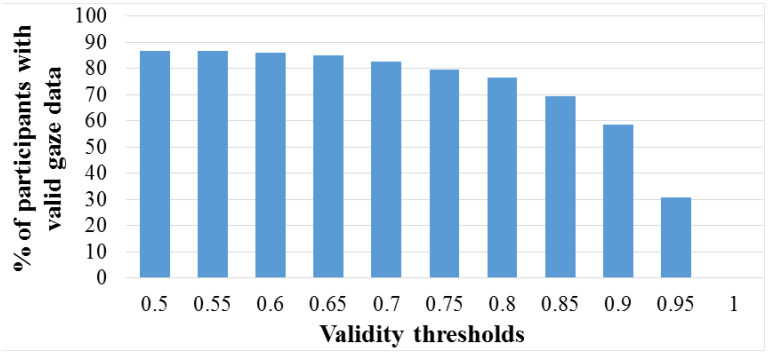

Figure 4.1 Number of participants with valid eye-tracking data as the strictness of validity threshold is increased from 0.5 to 1.

We study the tradeoff between data quality and amount available for training by comparing the accuracy of classifiers built on datasets with the following three validity thresholds:

- 0.9, which is the last threshold in Figure 4.1 that maintains high quality data without losing a large majority of participants (97 participants retained, i.e., 58% of all users); Similar thresholds have often been used in previous work on eye-tracking based user modeling, e.g., [Gingerich and Conati 2015; Jaques et al. 2014; Lallé et al. 2016a; Steichen et al. 2014];
- 0.6, which includes rather noisy data but a large pool of users (144 participants retained, i.e., 86% of all users);
- 0.8, which is a compromise between the two other thresholds (127 participants retained, i.e., 77% of all users).

Chapter 5: Classification Experiments and Results

We evaluate the prediction of our user characteristics (PS, SpM, VisScan, VisWM, VisLit) using a two-stage approach. The first stage ascertains whether we can build classifiers that can beat a majority-class baseline by evaluating two classification algorithms available in the caret package [Kuhn 2008] in R: Boosted logistic regression (LB); and Random Forest (RF). Classifier performance is measured by their accuracy (proportion of correct predictions). We focus on these algorithms because in previous work they produced good results for predicting various user states during visualization processing, e.g., [Steichen et al. 2014; Lallé et al. 2016b].

The second stage takes the best classifier identified in stage one, and investigates the impact of data quality on prediction accuracy, as well as how early during interaction with MQ we can obtain accurate predictions.

5.1 Feasibility of Predicting User Characteristics

In this first stage, we evaluate the performance of LB, RF, and majority-class baselines as predictors for the target binary labels. It should be noted that baselines are slightly different for datasets with different thresholds because they include different users. For all combinations of user characteristics (5), window lengths (10) and validity thresholds (3), LB, RF and the appropriate baselines were trained and evaluated in 10-fold cross validation over users, namely at each fold users in the test set do not appear in the training set. The process was repeated 25 times (runs) to strengthen the stability and reproducibility of the results. The accuracy of each classifier is averaged over the 10 folds and the 25 runs.

User Char	Threshold 0.6				Threshold 0.8				Threshold 0.9			
	Accuracy (%)			F-Statistic	Accuracy (%)			F-Statistic	Accuracy (%)			F-Statistic
	RF	LB	Baseline		RF	LB	Baseline		RF	LB	Baseline	
PS	**60.3**†	55.8†	50.8	$F_{2,7497}$=496.2	**60.6**†	56.3†	50.8	$F_{2,7497}$=523.436	**58.1**†	54.1†	51.5	$F_{2,7497}$=194.1
SpM	**60.9**†	58.2†	50.6	$F_{2,7497}$=587.08	**61.3**†	58.7†	50.4	$F_{2,7497}$=663.93	**61.2**†	57.5†	50.4	$F_{2,7497}$=504.99
VisScan	**57.3**†	55.4†	52.5	$F_{2,7497}$=125.469	**56.7**†	54.9†	51.8	$F_{2,7497}$=124.778	**56.2**†	53.7†	51.9	$F_{2,7497}$=82.02
VisWM	**57.5**†	54.7†	50.1	$F_{2,7497}$=282.09	**58.8**†	55.5†	50.5	$F_{2,7497}$=354.177	**60.3**†	55.7†	52.9	$F_{2,7497}$=225.75

Table 5.1 Accuracies (averaged across data windows) and F-statistics for the main effects of classification algorithm (RF, LF and baseline) for each of PS, SpM, VisScan and VisWM, and for each data validity threshold tested. † indicates that RF or LB significantly beat the corresponding baseline. Bold indicates that RF significantly beat LB.

To formally compare the obtained accuracies, for each of the five user characteristics, and for each of the three thresholds, we run a univariate General Linear Model (GLM) [Field 2012] with classification algorithm (3 levels) as factor and classification accuracy averaged across windows as the dependent measure[2]. Results show significant[3] main effects of classification algorithm for PS, SpM, VisWM, and VisScan, for all three thresholds (F statistics are reported in the "F-statistic" columns in Table 5.1). Pairwise comparisons for these main effects (Sidak correction applied to adjust for multiple comparisons[Field 2012]), indicate that LB always beats the baseline. RF always beats the baseline and also outperforms LB in all cases (see "Accuracy" in Table 5.1). We thus opt for RF as the algorithm to further investigate the effect of threshold in the second stage of analysis. Neither RF nor LB beat the baseline in predicting *VisLit*, thus this characteristic is dropped from the next stage (stage 2) of our analysis.

5.2 Effects of Data Validity Threshold

For each of the four user characteristics found to be predictable by RF in the previous subsection (Table 5.1), Figure 5.1 includes a graph showing, for the 3 validity thresholds, classifier accuracy

[2] We run a separate model for each threshold (as opposed to including threshold as factor in one model) because this approach is better to compare classifiers for each threshold with the baseline for that threshold.

[3] Statistical significance in this thesis is reported at $p < 0.05$.

over the 10 windows. To formally compare the accuracies with different thresholds, for each of the 4 user characteristics, we run a linear mixed-effects ANOVA [Field 2012] with validity threshold (3 levels) and window length (10 levels) as factors along with classification accuracy as the dependent variable. The last column of Table 5.2 reports the see F-statistics for these models.

There is a main effect of validity threshold for PS, VisScan and VisWM. The results of pairwise comparisons on validity threshold (Sidak correction applied) for these three characteristics are summarized in the second column of Table 5.2. Thresholds are ordered by prediction accuracy, e.g., "0.6 > 0.9" indicates RF with data validity at 0.6 performed better than RF with data validity at 0.9. Underlining indicates that the differences between thresholds underlined together are not statistically significant.

User Char.	**Ranking of accuracy at different threshold**	**Main Effect (F-Statistic)**
PS	0.8 > 0.6 > 0.9	$F_{2,7470}= 23.78^{\dagger}$
SpM	0.8 > 0.9 > 0.6	$F_{2,7470}= 1.02$
VisScan	0.6 > 0.8 > 0.9	$F_{2,7470}= 7.10^{\dagger}$
VisWM	0.9 > 0.8 > 0.6	$F_{2,7470}= 34.46^{\dagger}$

Table 5.2 Comparison of RF accuracy with different data validity thresholds, † indicates that RF significantly beat the corresponding baseline.

Table 5.2 shows that for PS, both 0.8 and 0.6 are the best thresholds (there is no significant difference between their accuracies). For VisScan, validity threshold 0.6 provides the best classification accuracies. For VisWM, the best validity threshold is 0.9. No main effect of validity threshold was found for SpM, meaning that there are no significant differences in classification accuracies among the three thresholds for this characteristic.

It is notable that for PS, SpM, and VisScan, a validity threshold of 0.6 is either the best or is tied for the best. Recall that a threshold of 0.6 allows for more eye-tracking data to be used for training our models, but the data is noisier. When 0.6 provides top predictive accuracies, it means that having more training data outweighs or compensates for the need for having clean data. These findings provide preliminary evidence that accurate predictions can be made for users with rather noisy data, which is what will likely be available to the classifier if these predictions are to be used to guide adaptive support to users of MQ in real-world settings. As Table 5.2 shows, VisWM is the only user characteristic for which having more data is less important than having clean data, possibly because the patterns indicative of this ability are subtler and thus more prone to be learned incorrectly by a classifier with noisier data. We will further discuss prediction of VisWM at the end of this section.

In the rest of this section, we focus on the classifier with the winning validity threshold for each user characteristic, and discuss when it achieves its maximum accuracy over the ten data windows representing the availability of classification data overtime during an interaction with MQ. For each classifier, we performed pairwise comparisons between its accuracy at each of the 10 windows and report in Table 5.3 (third column) the set of windows (ordered by accuracy) where classification accuracy was statistically equivalent and outperformed the accuracy in all other windows.

User Char	Optimal Validity Threshold	Best Window Lengths	Accuracy at Earliest Best Window
PS	.6	40>10>50	63.9%
SpM	.6	80>60>70>90	67.4%
VisScan	.6	30	64.0%
VisWM	.9	30	68.2%

Table 5.3 Windows for which the highest accuracies are achieved for the given validity thresholds.

The last column in Table 5.3 reports the accuracy at the earliest window reported in the previous column. For instance, for PS, windows 10, 40, and 50 have statistically equivalent accuracy and outperform all other windows. Thus, the earliest optimal prediction for PS can be obtained after only 10% of a user's interaction with the MQ task (only 15 seconds of interaction on average), with an accuracy close to 63.9%. Overall, Table 5.3 shows that the best predictions occur early not just for PS, but also for VisScan and VisWM (window 30). For SpM, the earliest best window is 60, i.e., slightly more than halfway in the task, which still leave substantial time to provide adaptation.

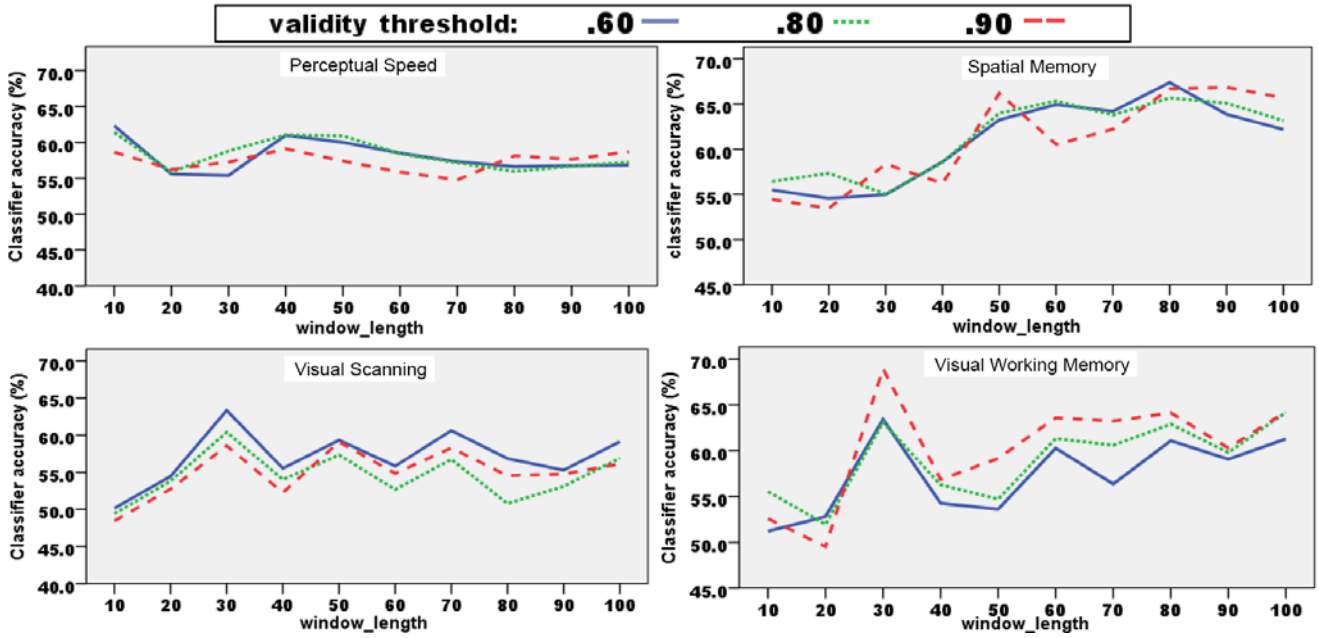

Figure 5.1 Accuracy of RF classifiers using data with three different validity thresholds for: PS, SpM, VisScan and VisWM.

5.3 Further Results for VisWM

The previous subsection showed that unlike the other user characteristics, the best predictions for VisWM are obtained using data with a high validity threshold of 0.9. Here, we investigate

whether a VisWM classifier trained using this high-quality data can still make good predictions on users with noisier (and thus more realistic) data. Specifically, we evaluated a RF classifier using 25-runs of 10-fold cross validation, where at each fold the training set included only users with data at the 0.9 validity threshold, and the test set included unseen users with data at the 0.6 validity threshold. A statistical analysis similar to the one in the previous subsection shows that this combined classifier and the VisWM classifier evaluated on data at the 0.9 threshold are not statistically different and both are significantly better than the model evaluated on data at the 0.6 threshold. In terms of actual accuracies and when they peak during interaction, the best (statically equivalent) windows for the combined classifier are 30, 80, and 70. The earliest of these windows (30), is the same as for the classifier with 0.9 validity and the accuracy at this window is 66.1%, which is comparable to the best accuracies for the other three user characteristics.

Chapter 6: Prediction using Action Features

In this chapter, we explore prediction of user characteristics using features derived from user interaction data collected during the user study described in Chapter 3. The data contains record of users performing various actions involving mouse clicks (e.g., to select the various tabs in screen 3, to rate a scenario etc.). Just like we did for eye-tracking data, we look at the actions performed only on screen 3. We divide them into two types: low-level and high-level actions.

6.1 Low-level action features

Low-level actions are generic actions related to mouse clicks and are not specific to any particular interface or task. Low-level action features are generated by EMDAT by calculating summary statistics on mouse clicks. Table 6.1, part a shows the list of all overall low-level features we considered for classification.

a) Overall low-level action Features (2):
Left click rate, Time to first left click
b) AOI low-level action Features for two out of the four AOIs (4):
Left click rate on AOI, Time to first left click on AOI

Table 6.1 Set of low-level Action features considered for classification.

By overall features we mean that these are generated for mouse clicks over the entire screen 3 instead of any defined AOIs. We included mouse click rate because it has been shown to be effective at predicting cognitive states in other interactive tasks [Lim et al. 2015] , thus we want to investigate whether this indicator might also inform classifiers for predicting our target user characteristics. We only consider rate of left mouse clicks because all mouse-related events in screen 3 are activated via the left button. Similarly, we included time to first click because it has

also shown potential for informing classifiers for user modeling, namely it has been used for early prediction of user failure with an interface [Liu et al. 2010]. Each of those features was generated for the interaction with the whole of screen 3. We also computed these features with respect to two of the AOIs that we defined in Chapter 4, namely the Map and the Legend AOI, because they are left clickable. Mouse clicks on AOIs have been used to understand the user's decision making process [Qin et al. 2013]. Furthermore, the use of AOI-related features from gaze data have been shown to have a strong impact on classification accuracy for user abilities related to the ones targeted here [Steichen et al. 2014]. Here, we want to ascertain whether there is merit in leveraging mouse-related features based on AOI to help predict our target user abilities.

6.2 High-level action features

High-level actions are actions that have significance for the interface in question. For our purposes, these are the actions that the users can perform to inform their ratings for various transit scenarios in screen 3 of MQ, i.e., the high-level actions that have significance from a decision-making perspective for MQ. Thus, we exclude actions that do not directly inform a user's choice (e.g., making optional additional comments, an option available for each scenario). The full list of high-level actions is in Table 6.2.

Switching tabs
Rating a scenario
Open instruction pop up in screen 3
Moving the map
Opening the legend
Zooming on the map

Table 6.2 Set of high-level actions

It should be noted that some of these high-level actions happen very infrequently or not at all for some users. For instance, 'Zooming on the map' was only done by 33 users [out of 166].

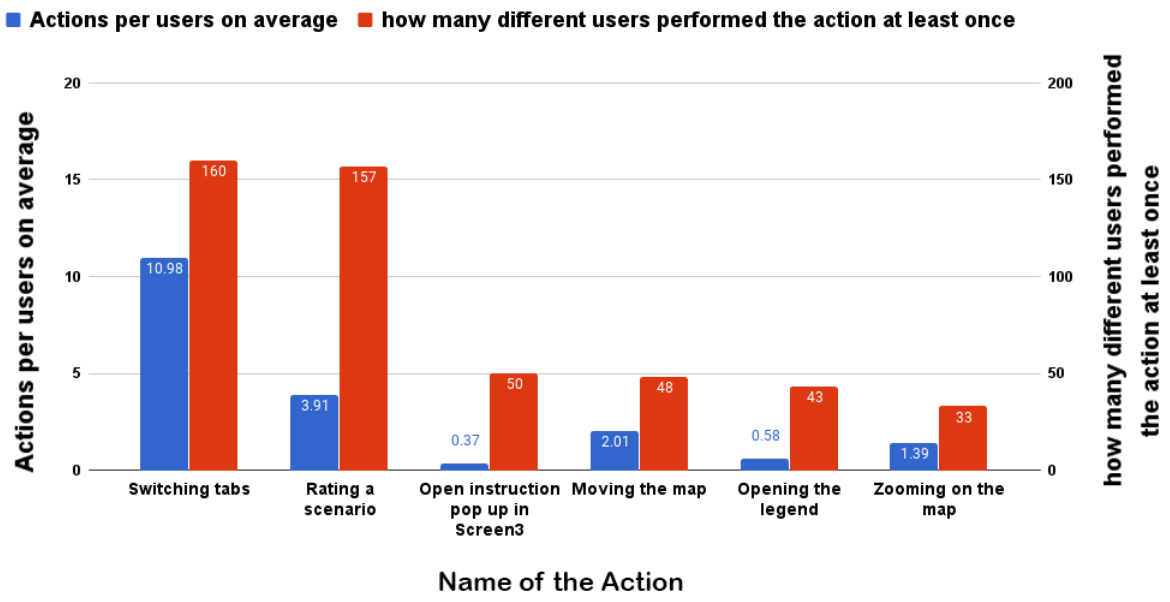

Figure 6.1 Average number of each type of high-level actions performed per user (blue bars) and number of different users who performed each action type at least once (red bars)

Figure 6.1 shows the number of time every high-level action has been performed per user (red bars), as well as how many different users have performed each action type at least once. From the figure, we can see that 'Open instruction pop up in Screen3' and 'Opening the legend' have the lowest frequencies of less than one, and were performed by less than a third of the users. Computing summary statistics such as mean and standard deviation over these infrequent actions is not meaningful. Instead, simply knowing whether a user performed these actions or not gives

the classification algorithms information that could be used for distinguishing between two users. Hence, for each of the two infrequent actions we included a binary feature symbolizing whether a participant performed that action or not, as listed in Table 6.3 (part a).

high-level action features (26)
a) Binary features for infrequent actions (2)
Open instruction pop up in Screen3', 'Opening the legend'
b) Features for the remaining high-level actions (24):
Frequency, Mean & Std. deviation for interval between an action and the subsequent one performed
Time before first action
Proportion of time spent on action, Proportion of actions performed

Table 6.3 Set of high-level Action features considered for classification.

- For the rest of the actions, we generated the features in Table 6.3 (part b), described below
- Time before first action indicates how much time the user has spent on screen 3 before performing a particular type of high-level action
- Frequency of an action indicates how many times that type of high-level action in performed during the time spent on screen 3.
- Time intervals between a given action and the subsequent one performed are derived from the logs and used to generate the following features:
  i. mean and standard deviation of those intervals which indicate, respectively, how much time, on average the user spent on a specific action before performing a new one, and how consistent this behavior was;
  ii. Proportion of overall time spent on an action by using the following formula

$$Proportion\ of\ time\ spent = \frac{Sum\ of\ all\ time\ intervals\ after\ that\ action}{Total\ time\ spent\ on\ Screen3}$$

- Proportion of actions performed, computed as the ratio between the number of times a certain high-level action has been performed divided by the total number of actions performed on screen3.

$$Proportion\ of\ actions\ performed = \frac{\#\ of\ times\ an\ action\ was\ performed}{Total\ \#\ of\ actions\ performed\ in\ Screen3}$$

Since, proportion of time and proportion of action features are both calculated for the entire screen, we calculate Pearson's correlation coefficient (r) [Field 2012] for each pair of features. None of them were highly correlated [r > 0.8] (see Appendix A.1- High Level Proportion Features). We also checked whether the low-level features were correlated with the generated high level action features. However, none of the features were highly correlated [see Appendix A.2- Low and High-Level Features]. Finally, as we did for eye-tracking data, we generated high and low-level action features for each of the 10 windows as reported previously; this enables us to ascertain how the accuracy of classifiers based on these features evolves with the amount of interaction data available.

6.3 Feasibility of Predicting User Characteristics using action features

As we did in Section 5.1 for eye-tracking-based classifiers, we start our analysis of action-based classifiers by finding the best classifier that beats a majority class baseline (if any) and evaluating how early during interaction with MQ we can obtain accurate predictions using only action features. We still consider only RF and LB as classification algorithms to maintain uniformity of analysis that was performed with eye-tracking data.

For all user characteristics (5) and window lengths (10), LB, RF and the appropriate baselines were trained and evaluated through 25 runs of 10-fold cross validation over users, as it was done with eye-tracking data. Figure 6.2 show the accuracy overtime for the three classifiers over the four user characteristics.

To formally compare the obtained accuracies, for each of the five user characteristics, we run a linear mixed-effects ANOVA [Field 2012] with classification algorithm (3 levels) and window length (10 levels) as factors, and with classification accuracy averaged across windows as the dependent variable.

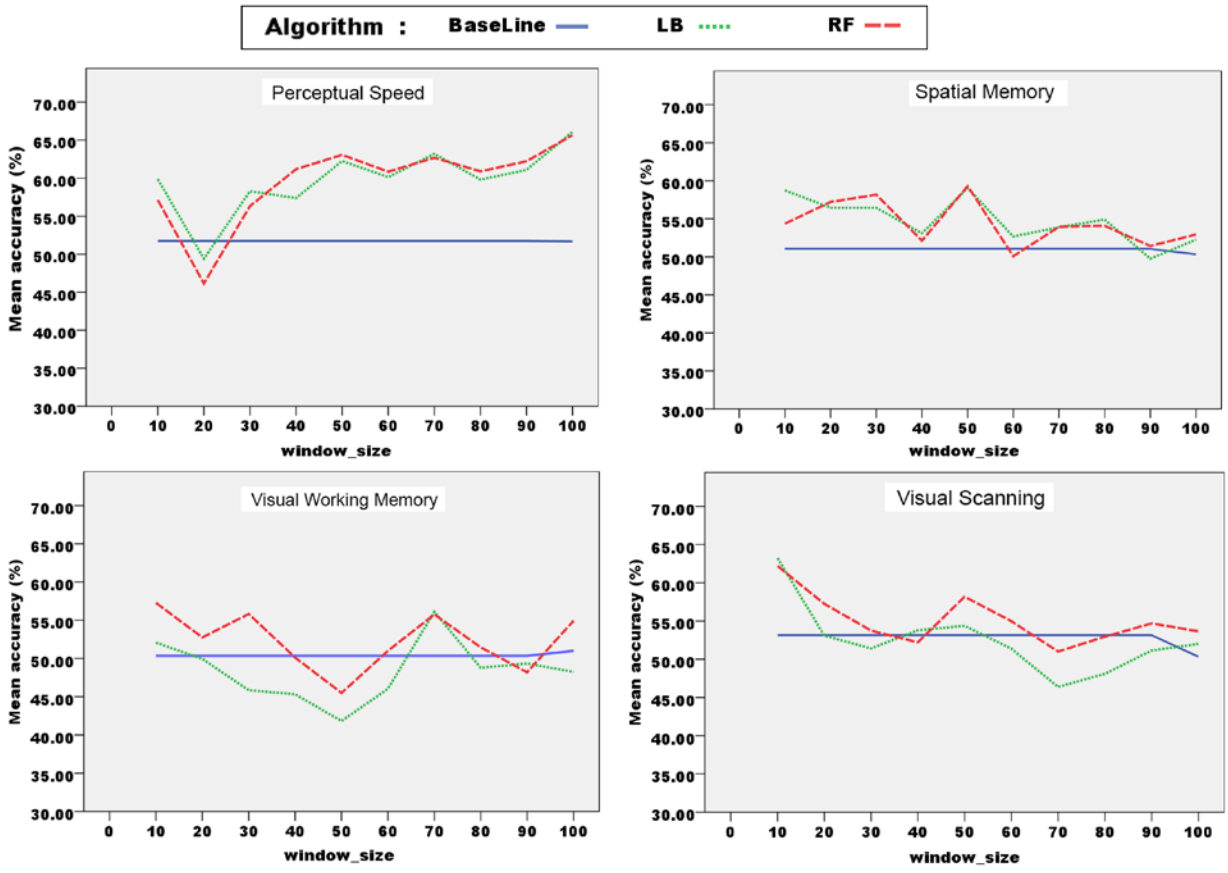

Figure 6.2 Results for PS, SpM, VisWM and VisScan with action features

Results show significant main effects of classification algorithm for PS, SpM, VisWM and VisScan, (F statistics are reported in the "F-statistic" column in Table 6.4). Pairwise comparisons for these main effects (Sidak corrections applied to adjust for multiple comparisons [Field 2012]), indicate that RF always beats the baseline; LB on the other hand fails to beat the baseline for VisWM and VisScan. Neither RF nor LB beat the baseline in predicting VisLit, which is consistent with the results we found for eye-tracking data. For PS and SpM, both classifiers beat the baseline, and there is no statistically significant difference in their accuracies. Thus, overall RF is the best performing classifier, because it consistently beats the baseline for PS, SpM, VisWM, and VisScan, and it is never outperformed by LB. Hence, we choose RF as the classifier for predicting user characteristics using action features.

User Char.	Ranking of accuracy of classifiers	Main Effect (F-Statistic)
PS	LB > RF > Baseline	$F_{2,7470} = 455.84^{\dagger}$
SpM	LB > RF > Baseline	$F_{2,7470} = 88.33^{\dagger}$
VisScan	RF > Baseline > LB	$F_{2,7470} = 42.29^{\dagger}$
VisWM	RF > Baseline > LB	$F_{2,7470} = 80.37^{\dagger}$

Table 6.4 Comparison of classification accuracies for the User characteristics that can be predicted better than baseline, † indicates that RF significantly beat the corresponding baseline.

Using the RF classifier, for each user characteristic we performed pairwise comparisons between classifier accuracy at each of the 10 windows, and report in Table 6.5 the best windows, ordered by accuracy, namely the set of windows where classification accuracy was statistically equivalent and outperformed (with statistical significance) the accuracy in all other windows. The last column in Table 6.5 reports the accuracy at the earliest window listed in the previous

column. For instance, for PS windows 100, 50, and 70 have statistically equivalent accuracy and outperform all other windows. Thus, the earliest optimal prediction for PS can be obtained after only 50% of a user's interaction with the MQ task, with an accuracy of 63%. Table 6.5, shows that for the four user characteristics, the best predictions using our approach occur no later than halfway in the task (i.e., window 50 for PS). Also, for VisScan only window length 10 is listed, since it outperforms every other window.

In the next chapter, we compare the performance of action data, eye-tracking data and the combination of the two for predicting our target user characteristics.

User Char	Chosen Classifier	Best Window Lengths	Accuracy at Earliest Best Window
PS	RF	100>50>70	63.06%
SpM	RF	50>30>20	57.2%
VisScan	RF	10	62.2%
VisWM	RF	10>30>70>100	57.3%

Table 6.4 Windows for which the highest accuracies are achieved for the chosen classifier (RF).

Chapter 7: Comparing the data sources

In Chapters 5 and 6, we have seen that both eye-tracking and Action features can be used to predict user characteristics. Eye-tracking and Action features can be considered as different data sources for prediction. In this section, we evaluate their effectiveness, i.e., we compare their prediction accuracies to see how good or bad they are in predicting user characteristics. We also introduce a new data source called combined by fusing eye-tracking and Action features together. For the combined datasets, we choose a threshold of 0.6 for the eye-tracking portion of the datasets. For 3 out of 4 user characteristics for which we beat the baseline [PS, VisScan and SpM], this is the threshold [0.6] that gave the best prediction accuracies when using eye-tracking data alone, thus it is natural to use this threshold for data fusion for the aforementioned 3 user characteristics. However, for VisWM, 0.9 had the best performance. The reasoning behind choosing 0.6 as the default threshold for VisWM is to build models with sufficient generalizability. A validity threshold of 0.6 corresponds to data which is rather noisy but also includes a sufficiently large pool of users, hence more general in nature. At this point, this is the same tradeoff we mention in Section 4.3. We address this tradeoff at this level by making the decision of having a larger pool of users for the fused data source instead of having cleaner but a smaller pool of user data.

We use the same approach as previous chapters by generating features in the combined dataset for each of the 10 interaction windows; in order to simulate increasing amounts of data available for classification as the interaction proceeds. We do not consider VisLit in this analysis, as neither action nor eye-tracking beat the baseline in predicting it, hence trying to predict VisLit using the combined data source does not hold much value.

7.1.1 Finding the Best Data Source and Feasibility of Early Predictions using Best Data Source

Using separate datasets for all 3 data sources [Action, Eye-Tracking and Combined], for all user characteristics (4) and window lengths (10), RF was trained and evaluated through 25 runs of 10-fold cross validation over users. As reported before, 0.6 was used as the default threshold for eye-tracking features in datasets associated with both eye-tracking and combined data sources. Figure 7.1 shows the graphs with average accuracies, for the 3 data sources over 10 windows. To formally compare the obtained accuracies, for each of the five user characteristics, we run a linear mixed-effects ANOVA [Field 2012] with data source (3 levels) and window length (10 levels) as factors along with classification accuracy averaged across windows as the dependent variable.

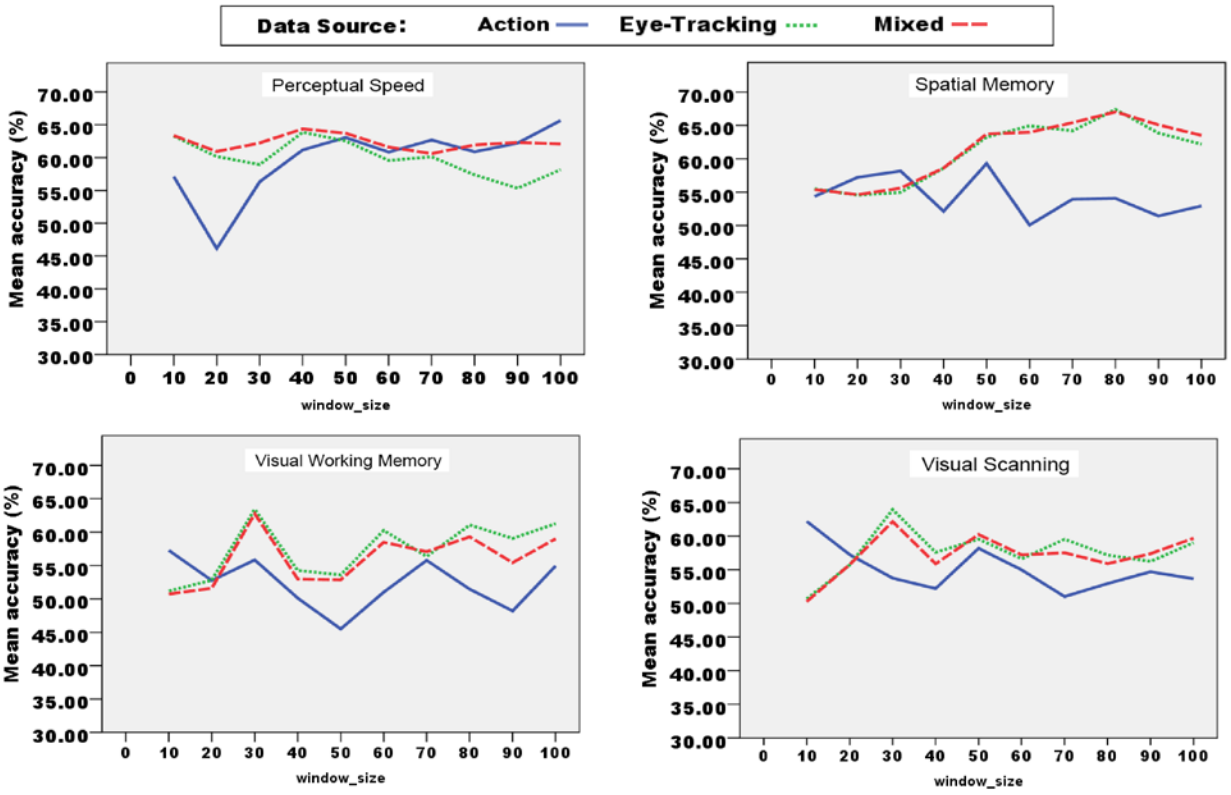

Figure 7.1 Comparison of accuracies for action, eye-tracking and mixed featuresets for PS, SpM, VisWM and VisScan

Results show significant main effects of data source for PS, SpM, VisWM and VisScan, (F statistics are reported in the "F-statistic" columns in Table 7.1). Pairwise comparisons for these main effects (Sidak correction applied to adjust for multiple comparisons [Field 2012]), indicate eye-tracking performs the best (with VisWM) or is tied for the best with Combined (with SpM, VisScan) in terms of prediction accuracies. For PS, combined performs significantly better than the other two data sources. Regardless of the best performing data source, the best performance requires eye-tracking data.

User Char.	Ranking of accuracy for various data source	Main Effect (F-Statistic)
PS	Combined > Eye-Tracking > Action	$F_{2,7470} = 37.26$
SpM	Combined > Eye-Tracking > Action	$F_{2,7470} = 254.24$
VisScan	Eye-Tracking> Combined > Action	$F_{2,7470} = 31.99$
VisWM	Eye-Tracking> Combined > Action	$F_{2,7470} = 116.69$

Table 7.1 Windows for which the highest accuracies are achieved for the chosen classifier (RF).

For each data source, we performed pairwise comparisons between its accuracy at each of the 10 windows, and report in Table 7.2 the best windows for the chosen data source reported previously, ordered by accuracy, namely the set of windows where classification accuracy was statistically equivalent and outperformed (with statistical significance) the accuracy in all other windows. The last column in Table 7.2 reports the accuracy at the earliest window reported in the previous column. For instance, for SpM, windows 80, 60, 70 and 90 have statistically equivalent accuracy and outperform all other windows. Thus, the earliest optimal prediction for SpM can be obtained slightly more than halfway in the task i.e., 60%. Table 7.2 shows that this is the latest window for optimal prediction: for the other three user characteristics, best prediction can be done as early as the 30% window (for VisScan and VisWM) or 10% (for PS).

Interestingly, for PS, all windows have statistically similar accuracies with the winning Combined data source. This is a result of the fact that, for this user characteristic, eye-tracking performs best in the first half of the interaction [40>10>50] as reported in Table 5.3 whereas action, as reported in Table 6.4, performs well in later windows [100>50>70] with almost similar average accuracies. This leads to the intuition that RF manages to pick the right features at each window with the combined data source datasets, i.e., presumably more eye-tracking features are selected in the earlier windows whereas in later windows RF most likely choose more action features. Thus, for PS, combined is the winning data source because it ensures steadier accuracy overtime, but it does not substantially improve the value of the peak accuracy (63.5%) compared to the other two data sources (63.9% for eye-tracking and 63% for actions).

In contrast to what happens with PS, for SpM, VisWM and VisScan, the action data source does better than the other two data sources earlier in the interaction, as seen in Figure 7.1, though it is never the best data source overall. For these user characteristics, the best data sources perform better in later windows. Thus, using action only gives us the opportunity to have early predictions, if getting early predictions is the priority, instead of getting the best possible results.

User Char	Chosen DataSource	Best Window Lengths	Accuracy at Earliest Best Window
PS	Combined	All windows have similar performance	63.4%
SpM	Eye-Tracking	80>60>70>90	64.95%
VisScan	Eye-Tracking	30	63.95%
VisWM	Eye-Tracking	30>100>80>60	63.42%

Table 7.2 Windows for which the highest accuracies are achieved for the chosen data source.

Chapter 8: Conclusion

In this thesis, we investigated if a user's cognitive abilities relevant for processing information visualizations can be predicted only either using eye-tracking or action data during interaction with MetroQuest (MQ), a visualization-based system designed to engage the public in environmental decision making. We also used data fusion to generate a third featureset called combined and predicted user cognitive abilities on the combined featureset. Then, we compared these data sources to find out which data source is the most suitable to predict each of the user characteristics considered.

We showed that a Random Forest classifier outperforms a majority-class baseline in predicting four of the five user cognitive abilities we tested: perceptual speed (PS), visual WM (VisWM), spatial memory (SpM), and visual scanning (VisScan) using eye-tracking data. Our results are important because previous findings on predicting user abilities during visualization processing were obtained for fictitious tasks done on bar and radar charts, whereas here we consider two different visualizations used for a task resembling how MQ is employed in real-life settings. Moreover, previous findings were obtained only for PS and VisWM, whereas we also showed the feasibility of predicting two additional cognitive abilities. Thus, our findings are a step toward showing the generality of predicting a variety of user cognitive abilities during visualization processing. These predictions are motivated by the long-term goal of devising user-adaptive visualizations that can recognize and adapt to the specific abilities of their users.

We also investigated how noise in eye-tracking data influences our prediction. For PS, SpM, and VisScan we found that training our classifiers with noisier but larger datasets worked better that

having cleaner but fewer data. These results suggest that further investigation on the impact of gaze data validity in user modeling is worthwhile, because it may eventually reduce the efforts researchers have to put in obtaining high validity data, at least for specific user modeling tasks. As for VisWM, the best accuracies were obtained with a classifier trained with high validity data. However, we showed that this classifier can still make good predictions on noisier data containing up to 40% invalid samples per user. Investigating prediction on noisy data is important to gauge the applicability of eye-tracking based user models in real-world settings.

Just as we did for eye-tracking data, we also showed that action data, which is more easily available than eye-tracking, can also be used to predict PS, VisWM, SpM and VisScan. We also fused these two data sources to yield similar results. Additionally, we showed that, while some previous work indicates that the fused featureset is best for predicting user's cognitive states, it is not always the case. In fact, for three out of the four user characteristics [VisWM, SpM and VisScan] for which we beat a majority class baseline, eye-tracking gives the best overall performance. For PS, combined gives the best overall performance; a featureset which contains eye-tracking data. This works as a proof of concept for leveraging eye-tracking data for predicting user's cognitive abilities, even in real world scenarios. On the other hand, we showed that while action data might not give the best results overall, it gives good predictions at the beginning of the interaction. Thus, it gives us the indication that while eye-tracking is the most suited for predicting user characteristics among the data sources leveraged, use of action data has merit if we want early predictions.

Thus, as part of future work for this research, we plan to continue experimenting with both eye-tracking and action data collected in realistic settings. One direction of research is to investigate the tradeoff between having early predictions using action data vs getting better accuracies using eye-tracking data. We also plan to investigate ways to increase the performance of our classification models. We can use feature selection and experiment with other classifiers in order to get better accuracies. The datasets used had a lot of features compared to number of data points. Thus, another potential direction of future investigation would be to use dimensionality reduction techniques such as PCA, which might yield better accuracies.

Bibliography

Jae-wook Ahn and Peter Brusilovsky. 2013. Adaptive visualization for exploratory information retrieval. *Information Processing Management* 49, 5 (September 2013), 1139–1164.

Roman Bednarik, Shahram Eivazi, and Hana Vrzakova. 2013. . In Yukiko I. Nakano, Cristina Conati, & Thomas Bader, eds. *Eye Gaze in Intelligent User Interfaces*. London: Springer London, 111–134.

Robert Bixler and Sidney D'Mello. 2015. Automatic Gaze-Based Detection of Mind Wandering with Metacognitive Awareness. In *Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization*. Dublin, Ireland: Springer, 31–43.

J. Boy, R.A. Rensink, E. Bertini, and J.D. Fekete. 2014. A Principled Way of Assessing Visualization Literacy. *IEEE Transactions on Visualizations and Computer Graphics* 20, 12 (décembre 2014), 1963–1972.

Cristina Conati, Giuseppe Carenini, Dereck Toker, and Sébastien Lallé. 2015. Towards User-Adaptive Information Visualization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Gitta O. Domik and Bernd Gutkauf. 1994. User modeling for adaptive visualization systems. In *Proceedings of the conference on Visualization'94*. IEEE Computer Society Press, 217–223.

Ruth B. Ekstrom, John W. French, Harry H. Harman, and Diran Dermen. 1976. *Manual for kit of factor referenced cognitive tests*, Educational Testing Service Princeton, NJ.

Andy Field. 2012. *Discovering Statistics Using IBM SPSS Statistics* 4. ed., London: Sage Publications Ltd.

Keisuke Fukuda and Edward K. Vogel. 2009. Human Variation in Overriding Attentional Capture. *J. Neurosci.* 29, 27 (July 2009), 8726–8733.

Matthew Junghyun Gingerich and Cristina Conati. 2015. Constructing Models of User and Task Characteristics from Eye Gaze Data for User-Adaptive Information Highlighting. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

David Gotz and Zhen Wen. 2009. Behavior-driven visualization recommendation. In *Proceedings of the 14th International conference on Intelligent user interfaces*. New York, NY, USA: ACM, 315–324.

Kenneth Holmqvist ed. 2011. *Eye tracking: a comprehensive guide to methods and measures*, Oxford ; New York: Oxford University Press.

Dandan Huang et al. 2015. Personal visualization and personal visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 21, 3 (2015), 420–433.

Shamsi T. Iqbal, Piotr D. Adamczyk, Xianjun Sam Zheng, and Brian P. Bailey. 2005. Towards an index of opportunity: understanding changes in mental workload during task execution. In ACM Press, 311.

Young-Min Jang, Rammohan Mallipeddi, and Minho Lee. 2014. Identification of human implicit visual search intention based on eye movement and pupillary analysis. *User Modeling and User-Adapted Interaction* 24, 4 (October 2014), 315–344.

Natasha Jaques, Cristina Conati, Jason M. Harley, and Roger Azevedo. 2014. Predicting Affect from Gaze Data during Interaction with an Intelligent Tutoring System. In Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, & Kitty Panourgia, eds.

Proceedings of the 12th International Conference on Intelligent Tutoring Systems. Lecture Notes in Computer Science. Honolulu, HI, USA: Springer, 29–38.

Samad Kardan and Cristina Conati. 2013. Comparing and Combining Eye Gaze and Interface Actions for Determining User Learning with an Interactive Simulation. In *Proceedings of UMAP, 21st International Conference on User Modeling, Adaptation and Personalization*.

Samad Kardan and Cristina Conati. 2012. Exploring gaze data for determining user learning with an interactive simulation. In *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization*. Montreal, QC, Canada: Springer, 126–138.

M. Kuhn. 2008. Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 5 (2008), 1–26.

Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2016a. Predicting confusion in information visualization from eye tracking and interaction data. In *Proceedings on the 25th International Joint Conference on Artificial Intelligence*. 2529–2535.

Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2016b. Prediction of individual learning curves across information visualizations. *User Modeling and User-Adapted Interaction* 26, 4 (2016), 307–345.

Yee Mei Lim, Aladdin Ayesh, and Martin Stacey. 2015. Using Mouse and Keyboard Dynamics to Detect Cognitive Stress During Mental Arithmetic. In *Intelligent Systems in Science and Information 2014*. Springer, 335–350.

Chao Liu, Ryen W. White, and Susan Dumais. 2010. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR conference on Research and development in information retrieval*. ACM, 379–386.

Mohamed Mouine and Guy Lapalme. 2012. Using Clustering to Personalize Visualization. In *Proceedings of the 16th International Conference on Information Visualisation*. Montpellier, France: IEEE, 258–263.

Kasia Muldner, Winslow Burleson, and Kurt VanLehn. 2010. “Yes!”: using tutor and sensor data to predict moments of delight during instructional activities. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 159–170.

Kawa Nazemi, Wilhelm Retz, Jörn Kohlhammer, and Arjan Kuijper. 2014. User similarity and deviation analysis for adaptive visualizations. In *Human Interface and the Management of Information. Information and Knowledge Design and Evaluation*. Springer, 64–75.

Kristien Ooms, Philippe De Maeyer, and Veerle Fack. 2014. Study of the attentive behavior of novice and expert map users using eye tracking. *Cartogr. Geogr. Information Science* 41, 1 (2014), 37–54.

André Pimenta, Davide Carneiro, Paulo Novais, and José Neves. 2013. Monitoring mental fatigue through the analysis of keyboard and mouse interaction patterns. In *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 222–231.

Linchan Qin, Ning Zhong, Shengfu Lu, and Mi Li. 2013. Decision prediction using visual patterns. *Fundam. Informaticae* 127, 1–4 (2013), 545–560.

S. Lallé, C. Conati, and G. Carenini. 2017. Impact of Individual Differences on User Experience with a Real-World Visualization Interface for Public Engagement. In *International Conference on User Modeling, Adaptation, and Personalization*.

Oleg Špakov. 2012. Comparison of eye movement filters used in HCI. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 281–284.

Ben Steichen, Cristina Conati, and Giuseppe Carenini. 2014. Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *ACM Transactions on Interactive Intelligent Systems* (2014).

Dereck Toker, Cristina Conati, Giuseppe Carenini, and Mona Haraty. 2012. Towards adaptive information visualization: on the influence of user characteristics. In *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization*. UMAP'12. Berlin, Heidelberg: Springer-Verlag, 274–285.

Dereck Toker, Cristina Conati, Ben Steichen, and Giuseppe Carenini. 2013. Individual user characteristics and information visualization: connecting the dots through eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York, NY, USA: ACM, 295–304.

Anatoly Yelizarov and Dennis Gamayunov. 2014. Adaptive Visualization Interface That Manages User's Cognitive Load Based on Interaction Characteristics. In *Proceedings of the 7th International Symposium on Visual Information Communication and Interaction*. ACM, 1.

Caroline Ziemkiewicz, R. Jordan Crouser, Ashley Rye Yauilla, Sara L. Su, William Ribarsky, and Remeo Chang. 2011. How locus of control influences compatibility with visualization style. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*. Providence, RI, USA: IEEE, 81–90.

Appendices

Appendix A - Correlation Analysis

As reported in Chapter 6 we calculated Pearson's correlation coefficient [Field 2012] for proportion of time and proportion of action features [see Table 6.3 for details]. We calculated the same to see whether there is correlation between high level and low-level action features [see Table 6.1]. Here, we describe our findings in detail.

A.1 High Level Proportion Features

In 6.2 we defined proportion of overall time spent as the ratio between as sum of all time intervals after that action divided by the total time spent on screen 3. The formula for calculating this feature is as follows:

$$Proportion\ of\ time\ spent = \frac{Sum\ of\ all\ time\ intervals\ after\ that\ action}{Total\ time\ spent\ on\ Screen3}$$

and Proportion of actions performed, computed as the ratio between the number of times a certain high-level action has been performed divided by the total number of actions performed on screen3. The formula for calculating this feature is as follows:

$$Proportion\ of\ actions\ performed = \frac{\#\ of\ times\ an\ action\ was\ performed}{Total\ \#\ of\ actions\ performed\ in\ Screen3}$$

Here it should be noted that for four out of six high level actions we calculated the features reported above. For the rest two ['Opening the legend' and 'Zooming on the map'], we do not calculate these two features because they are infrequent actions [see Section 6.2 for details]. For the rest of the four more frequent actions, we calculated the Pearson's correlation coefficient for

each pair of proportion of time spent and proportion of actions performed features. The results are reported in Table A.1.1.

Feature name	Pearson's coefficient value	Correlation type
Proportion of time spent for 'switching tabs'	0.64	Moderately strong positive correlation
Proportion of actions performed for 'switching tabs'		
Proportion of time spent for 'Rating a scenario'	0.6	Moderately strong positive correlation
Proportion of actions performed 'Rating a scenario'		
Proportion of time spent for 'Open instruction pop up'	0.2	Weak positive correlation
Proportion of actions performed 'Open instruction pop up'		
Proportion of time spent for 'Moving the map'	0.57	Moderately strong positive correlation
Proportion of actions performed 'Moving the map'		

Table A.1.1 Pearson's correlation coefficient values for proportion of time and proportion of actions features

A.2 Low and High-Level Features

We calculated Pearson's correlation coefficient between low-level features and high level-features. We report the correlation coefficient among high-level frequency features with other low-level features in Table A.1.2. It should be noted that only the ones where there has been some amount of correlation among the low-level features and high level frequency features are reported.

High-level action feature name		Pearson's coefficient value	Correlation type
Frequency Switching Tabs			
Low Level Features	Left click rate	.47	weak positive correlation
	Time to first left click	-.18	weak negative correlation
	Left click rate on 'Map'	.26	weak positive correlation
	Left click rate on 'Description of the transit scenario'	.36	weak positive correlation
	Time to first left click on 'Description of the transit scenario'	.02	weak positive correlation
	Left click rate on 'Legend of the map'	.32	weak positive correlation
	Time to first left click on 'Legend of the map'	.18	weak positive correlation
Frequency Moving the map			
Low Level Features	Left click rate	.32	weak positive correlation
	Time to first left click	-.20	weak negative correlation
	Left click rate on 'Map'	.52	weak positive correlation
	Left click rate on 'Description of the transit scenario'	.33	weak positive correlation
	Time to first left click on 'Description of the transit scenario'	.37	weak positive correlation
	Time to first left click on 'Legend of the map'	.26	weak positive correlation

High-level action feature name		Pearson's coefficient value	Correlation type
Frequency Zooming on the map			
Low Level Features	Left click rate	.33	weak positive correlation
	Time to first left click	-.18	weak negative correlation
	Left click rate on 'Map'	.26	weak positive correlation
	Left click rate on 'Description of the transit scenario'	.36	weak positive correlation
	Time to first left click on 'Description of the transit scenario'	.02	weak positive correlation
	Left click rate on 'Legend of the map'	.32	weak positive correlation
	Time to first left click on 'Legend of the map'	.18	weak positive correlation
Frequency Rating a scenario			
Low Level Features	Left click rate	.36	weak positive correlation
	Time to first left click	-.20	weak negative correlation
	Left click rate on 'Deviation Chart	.17	weak positive correlation

Table A.1.2 Pearson's correlation coefficient values for high level frequency and low-level features