# Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data

**Sébastien Lallé, Cristina Conati, Giuseppe Carenini**
The University of British Columbia, Vancouver B.C., Canada
{lalles, conati, carenini}@cs.ubc.ca

## Abstract

Confusion has been found to hinder user experience with visualizations. If confusion could be predicted and resolved in real time, user experience and satisfaction would greatly improve. In this paper, we focus on predicting occurrences of confusion during the interaction with a visualization using eye tracking and mouse data. The data was collected during a user study with ValueChart, an interactive visualization to support preferential choices. We report very promising results based on Random Forest classifiers.

## 1 Introduction

Confusion has been found to hinder user experience and satisfaction with user interfaces [e.g., Nadkarni and Gupta 2007] and Information Visualization (InfoVis) [e.g., Lee *et al.* 2016; Yi 2008]. To date, most work on developing methods to detect and prevent or resolve confusion in real-time during interaction has been in the field of Intelligent Tutoring Systems (ITS), [Bosch *et al.* 2015; Baker *et al.* 2012; D'Mello and Graesser 2007]. Building on this work, we investigate how to predict confusion during the interaction with a visualization-based interface, with the long-term goal of devising intelligent user-adaptive visualizations that can provide personalized interventions to help confused users. 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].

In this paper, we investigate machine learning models to predict occurrences of confusion during the interaction with ValueChart, an interactive visualization to support multi-criteria preferential choice. To make such predictions in real-time, we leverage both interaction data as well as eye tracking capturing users' gaze patterns, pupil size and head distance to the screen. Interaction data have been used before to predict confusion in computer games [Pentel 2015] and ITS [Baker *et al.* 2012]. However, we are the first to study how eye tracking can be used to predict confusion, with the rationale that eye tracking should be particularly informative in predicting confusion during visualization processing, as visual scanning is a fundamental component of working with a visualization. Furthermore, eye tracking has been shown to be a good predictor of affective states in educational systems [Jaques *et al.* 2014; Muldner *et al.* 2010] as well as user characteristics in InfoVis, [e.g., Steichen *et al.* 2014; Jang *et al.* 2014].

This work has two main contributions. The first is a proof of concept that confusion can be predicted in real-time in InfoVis, with 61% accuracy (significantly better than chance) for occurrences of confusion and a false positive rate of only 7.4%. The second contribution is evidence of the importance of eye tracking for building predictors of confusion in InfoVis, with the most informative sources of information being differences in the user's attention to the labels of the InfoVis, as well as variations in user pupil size and head distance to the screen.

## 2 Related Work

Previous work suggests that confusion can negatively impact user experience or satisfaction with an interface, e.g., [Rickenberg and Reeves 2000; Nadkarni and Gupta 2007]. In InfoVis, confusion has been linked to lower user performance and satisfaction in completing decision making tasks [Yi 2008], and was shown to affect novice users when they process an unfamiliar visualization [Lee *et al.* 2016].

In the field of Intelligent Tutoring Systems, predictors of confusion have been built by leveraging facial expressions [Bosch *et al.* 2015; D'Mello and Graesser 2007], posture [D'Mello and Graesser 2007], or students' interface actions and studying behavior [Baker *et al.* 2012]. In HCI, Pentel [Pentel 2015] leveraged mouse usage information to predict occurrences of confusion in a simple computer game. We extend this work by showing the feasibility of predicting confusion in real time in InfoVis, and we extend previous work on confusion prediction in general by using a new sensor, eye tracking.

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]. In InfoVis, gaze and pupil data have been investigated to predict long-term user traits (e.g., perceptual speed, visual and verbal working memory), as well as short-term properties such as task completion time, learning curve, or intention for visual search [Steichen *et al.* 2014; Lallé *et al.* 2015; Jang *et al.* 2014]. Still in InfoVis, [Yelizarov and Gamayunov 2014] tracked mouse and keyboard events to predict users' level of cognitive load and adjust the amount of information to be displayed accordingly. Their results showed that their adaptations positively impacted users' performance.

## 3 User Study

**ValueCharts.** Complex decisions can often be framed as preferential choices, i.e., the process of selecting the best option out of a set of alternatives characterized by a variety of attributes (e.g., select a car to buy, a university to attend, etc.). The dataset[1] used in this paper was collected from a user study using ValueChart, an interactive visualization for preferential choice [Conati *et al.* 2014]. Figure 1 shows an example of ValueChart for selecting rental properties among ten available alternatives (listed in the leftmost column), based on a set of relevant attributes (e.g., location, appliances, etc.). These attributes are arranged hierarchically in the top part of the ValueChart, with a column for each attribute in the central part of the display. The width of each column indicates the relative weight assigned to the corresponding attribute. The available alternatives (i.e., rental homes) are represented as the rows in the display. Each cell specifies how the alternative in that row fares with respect to the attribute in that column, indicated by the amount of filled cell color. In the rightmost part of the ValueChart, all values for each alternative are accumulated and presented as stacked bars, displaying the overall value of each alternative (e.g., *home4* is the best home in the example in terms of overall value).

The interactive functionalities available to support the decision process include: *(i)* inspecting the specific domain value of each attribute (e.g., the rent of home1 being equal to $500), by left clicking on the related alternative; *(ii)* sorting the alternatives with respect to a specific attribute (by double-clicking on the attribute name); *(iii)* swapping attribute columns (initiated via a left click on one of the attributes); and *(iv)* resizing the width of an attribute's column to see how that would impact the decision outcome (initiated via a left click on the column edge).

ValueChart has been extensively evaluated for usability and adopted in several applications (e.g., [Yi 2008; Wongsuphasawat *et al.* 2012]). It has, however, inherent complexity due to the nature of the task, which can still generate confusion in some users [Yi 2008; Conati 2013].

**Procedure.** 136 participants (age range 16 to 40, 75 female) were recruited from various departments at our university to perform 5 different types of tasks with ValueChart (e.g., retrieve the cheapest home or select the best home based on the aggregation of price and size). After 10 min of training with ValueChart, each participant repeated each task type 8 times in a randomized fashion to account for within-user variability, for a total of 40 tasks. This results in a total of 5440 trials (136 users × 40 trials).

While performing the tasks, the user's gaze was tracked with a Tobii T120, a non-intrusive eye-tracker embedded in the study computer monitor. In order to avoid possible confounds on pupil size due to lighting changes, the study was administered in a windowless room with uniform lighting. To compensate for physiological differences in pupil size among users, pupil diameter baselines were collected for each user by having them stare at a blank screen for ten seconds at the beginning of the study.

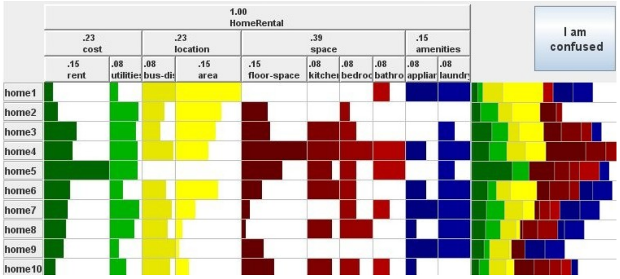


*Figure 1: An example of the main elements of ValueChart.*

**Collecting data on user confusion.** Collecting ground truth labels is one of the main challenges for building user models that can predict transient user states in an adaptive interface (e.g., [Porayska-Pomsta *et al.* 2013]), and a variety of methods have been proposed in the literature to address this challenge (see [Conati *et al.* 2013] for an overview).

After careful consideration of various options, we chose to have users self-report their confusion by clicking on a button labeled "I am confused" (see Fig. 1, top right). Users were instructed to use the button as follows (rephrased for brevity): "*[you should click the confusion button] if you feel that you want to ask the experimenter a question about something; if you are confused about the interface; if you are confused about the wording of a question. ...These are just a few examples, to show that confusion can occur in many unforeseeable ways, [which are all] OK reasons to click the confusion button.*" Participants were told that clicking on the button would have no effect on the interaction, it was just included for data collection. At the end of the study, each participant was shown replays of interaction segments centered around their reports of confusion, to verify that the report was intended and elicit the reason of the confusion. This collection method was evaluated via pilot studies before being deployed in the main experiment [Conati *et al.* 2013].

Confusion was reported in 112 trials (2% of all trials), with 80 users (59%) reporting confusion at least once during the study, with an average of 1.4 clicks (SD=1.9). There was never more than one click per trial. Reasons for confusion reported by participants include not understanding the tasks, perceived ambiguity in the textual or visual components of

[1] Data available at *http://www.cs.ubc.ca/~lalles/IJCAI16.html*

the interface, interactive functionalities not working as expected, alleged missing functionalities to solve the task.

To investigate the impact of confusion on users' performance, we ran an MANOVA with *task accuracy* and *task completion time* as the dependant variables, and the *occurrence of confusion* (2 levels: YES or NO) and *types of task* (5 levels) as factors. The MANOVA revealed a significant main effect of confusion on both accuracy ($F_{1,2159}$=108, $p$ < .001, $\eta^2_p$=.02) and completion time ($F_{1,2159}$=125, $p$ < .001, $\eta^2_p$=.02), with confusion trials having a lower accuracy and being longer than no-confusion trials. These results confirm the negative impact of confusion on performance reported in [Yi 2008] for a visualization based on a variation of ValueChart. It is notable that the impact exists even with the relatively low number of confusion reports in our dataset.

## 4 Predicting Confusion in ValueChart

We label trials as "*confusion*" when the user pressed the confusion button at least once during the trial, or "*no-confusion*" otherwise. Thus, predicting confusion in our dataset is a binary classification task: classify each trial as one in which the user might be experiencing confusion, or not. We compare a variety of features sets for this classification task (Section 4.2). In Section 4.1, we describe the datasets we generated to compute these feature sets.

### 4.1 Data Windows

To simulate the real-time prediction of confusion episodes, we use only users' data prior to the click on the confusion button. As there is no such click in no-confusion trials, we randomly generate a "pivot point" in each of those trials. In order to ascertain how much data leading up to an episode of confusion is needed to predict it, we built our feature sets (described below) by using two different windows of data: a *short window* captures data 5 seconds immediately before a confusion click[2] (or pivot point); a *full window* captures the whole episode of confusion by including data from a click (or pivot point) back to the beginning of the trial[3]. Full windows were on average 13.7s in length (SD=11.3s).

### 5.2 Predictive features

The eye tracking data collected during the study provide information on user gaze patterns (*Gaze*, from now), on changes in a user's pupil width (*Pupil*), and on the distance of the user's head from the screen (*Head Distance,* defined as the averaged distance of each eye to the screen), which are all good candidates to explore as predictors of confusion. Confusion is likely to impact how a user attends to elements of the visualization (Gaze), which has already been successfully used to predict user performance and other abilities with visualizations [Steichen *et al.* 2014; Lallé *et al.* 2015; Nazemi *et al.* 2014]. Pupil size has been associated to cognitive load [e.g., Granholm and Steinhauer 2004], which might be affected when the user experiences confusion. Pupil has also been shown to be a predictor of other mental states, such as mind wandering [e.g., Bixler and D'Mello 2015]. Head distance provides a rough indication of user posture, which has been shown to be a predictor of user engagement with a task [D'Mello and Graesser 2007]. Since ValueChart is interactive, we also leverage data on *mouse events*, which have been shown to predict user confusion during interaction with a computer game [Pentel 2015]. From all these measures (Gaze, Pupil, Head Distance, and Mouse Events) we derive four groups of features (listed in Table 1) that we evaluate in terms of their ability to predict user confusion with ValueChart.

| **a) Gaze Features (149)** |
|---|
| *Overall Gaze Features (9)* |
| Fixation rate |
| Mean & Std. deviation of fixation durations |
| Mean & Std. deviation of saccade length |
| Mean & Std. deviation of relative saccade angles |
| Mean & Std. deviation of absolute saccade angles |
| *AOI Gaze Features for each AOI (140)* |
| 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 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 |
| **c) Mouse Event Features (Overall and for each AOI) (32)** |
| Left click rate, Double click rate |
| Time to first left click, Time to first double click |

*Table 1: Sets of feature considered for classification.*

**Gaze features** (Table 1a) describe user's gaze patterns in terms of *fixations* (gaze maintained at one point on the screen), and *saccades* (quick movement of gaze from one fixation point to another). From this raw gaze data, we generated features that capture overall gaze activity on the screen, as well as activity over specific Areas of Interest (AOI), shown in Figure 2. We selected these features because they have been extensively used in HCI to capture differences in users' attention patterns over an interface [e.g., Holmqvist *et al.* 2015].

**Pupil and Head Distance features** (Table 1b) are generated by averaging the corresponding measures for each eye. Pupil size is adjusted using the pupil baseline collected during the study to get the *percentage change in pupil size* (PCPS), as defined in [Iqbal *et al.* 2005]. Features like the *mean*, *min*, *max* and *std.dev* are standard ways to measure fluctuations in a measure of interest, and have been used with pupil size to reveal individual differences while users work with an interface [Holmqvist *et al.* 2015] and with head distance to predict user boredom [Jaques *et al.* 2014]. We also included pupil size & head distance of the first and

[2] Windows actually end 1 second before a confusion click to avoid confounds associated with the specific intent to report confusion (e.g., fast straight saccades toward the button).

[3] Note that this window cannot include another confusion click because there is only one per trial in our dataset. Otherwise, full windows would go back only to the last episode of confusion.

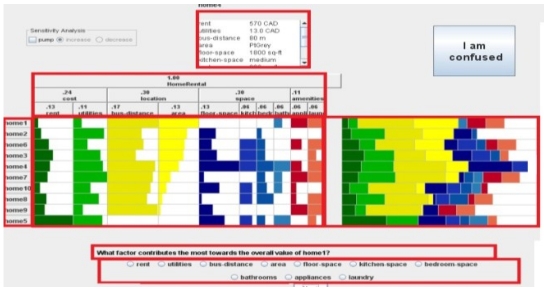
last fixation in the data window (see Section 4.1) as a way to capture variations of the measures between the closest and farthest data-points to the confusion click in that window.

**Mouse Event features** (Table 1c) are built upon summary statistics on left and double mouse clicks (the only two with associated functionalities), measured both over the whole screen and for the 7 specific AOIs defined in Figure 2. Mouse click rate has been shown to be effective at predicting cognitive states in other interactive tasks [Lim *et al.* 2015]. The time to first click has been used for early prediction of user failure with an interface [Liu *et al.* 2010], thus it might also be relevant to predict confusion because confused users are likely to make more errors.

From all the features described above, we derive a total of 8 feature sets that we will compare to predict confusion:

- One feature set for each of Gaze (G), Pupil (P), Head Distance (HD), and Mouse Events (ME) features.
- All features together (ALL).
- Only features derived from eye tracking data (G+P+HD), as these can be used in non-interactive InfoVis as well.
- Head distance and Pupil (P+HD) as these features are entirely independent from the layout of the visualization.
- Head Distance+MouseEvent (HD+ME), as the expensive eye tracker is not needed to collect these features (a webcam can reliably infer the distance to the screen).

We used each of the two data windows described in Section 4.1 to generate each of the 8 features sets above, for a total of 16 combinations.



rent 570 CAD
utilities 13.0 CAD
bus-distance 80 m
area PtGrey
floor-space 1800 sq-ft
kitchen-space medium

What factor contributes the most towards the overall value of home1?

○ rent ○ utilities ○ bus-distance ○ area ○ floor-space ○ kitchen-space ○ bedroom-space ○ bathrooms ○ appliances ○ laundry

*Figure 2: Areas of Interest (AOI) defined for ValueChart.*

## 4.3 Addressing data imbalance

Our dataset is highly imbalanced, with only 2% of confusion trials. Building meaningful classifiers on imbalanced dataset is challenging: a majority class classifier that always predicts no confusion would have a very high accuracy, but would be useless. Thus, to train our predictors of confusion, we balanced training data by using the well-known SMOTE algorithm for *data over-sampling* [Chawla *et al.* 2002]. Specifically, using SMOTE we generated "synthetic" confusion trials based on *k* nearest neighbors in the minority class (we used the default value *k=5*). To do so, SMOTE randomly duplicates a confusion trial (*ct*) and modifies it by sampling for each feature a new value along the line between c*t* and its *k* nearest neighbors. Next, no-confusion trials are randomly discarded (*down-sampled*) until the dataset is balanced. In our study, we over-sampled confusion trials by 200% (i.e., number of confusion trials is doubled) and 500%. These percentages appeared to be among the best ones when applied to similar imbalanced datasets in [Chawla *et al.* 2002], where they also show that more over-sampling is pointless.

## 4.4 Machine learning set up

To build our classifiers, we use Random Forest tuned with 100 trees using the Caret package in R [Kuhn 2008]. We chose Random Forest because previous work has shown that this learning algorithm performed well on similar prediction tasks [e.g., Pentel 2015; Wu 2015]. The classifiers are trained for all combinations of features sets (8), data window lengths (2) and SMOTE configurations (2) for a total of 32 classifiers. These classifiers are trained and evaluated with a process of 20-runs-10-folds nested cross-validation, which includes two levels (*inner* and *outer*) of cross-validation.

At the outer level the following process is repeated 20 times (runs) to strengthen the stability and reproducibility of the results. Data is randomly partitioned in 10 folds; in turn, each of the 10 folds is selected as a test set; the remaining 9 folds are SMOTE-balanced and used to train a classifier which is then tested on the test fold. The performance of each classifier is averaged across the outer test sets, and then again over the 20 runs. It should be noted that test sets at the outer level are not altered by SMOTE in any way.

We performed cross-validation over users, meaning that in each cross-validation fold, all trials of a given users are either in the training or in the test set. We kept in the folds a distribution of confusion data points similar to that in the whole dataset. As the outer test folds are strongly imbalanced, model performance is measured via sensitivity and specificity, two suitable measures when data are skewed [Kotsiantis *et al.* 2006], defined as follow:

- *Sensitivity* (or true positive rate): proportion of confusion trials that are correctly identified as such. It indicates the ability of the model to predict occurrences of confusion.
- *Specificity* (or true negative rate): proportion of no-confusion trials that are correctly identified as such. It indicates the ability of the model to avoid false positives (as 1 − *Specificity* is the proportion of false positives).

At the inner level of the nested cross-validation, Correlation Feature Selection [Kuhn 2008] is applied to remove highly correlated features. Next, the best *decision threshold* is selected for each Random forest classifier using ROC curves plotted on inner train data only. The threshold selected is the one that gives the best trade-off between sensitivity and (1 – specificity) by taking the closest point of the ROC curve to the point (0, 1) representing perfect classification [Fawcett 2006].

# 5 Results

We analyze the performance of the 32 classifiers defined in Section 4.4 by running a 3-way MANOVA with:

- *sensitivity* and *specificity* of the classifiers as the two dependent variables,

- *feature set* (8 levels), *window length* (2 levels) and *SMOTE configuration* (2 levels) as the factors.

The MANOVA reveals a main effect of both *feature set* ($F_{8,67}$=18.65, *p* < .001, $\eta^2_p$=.59) and *SMOTE configuration* ($F_{1,26}$=9.58, *p* = .005, $\eta^2_p$=.17). No main effect of data *window length*, nor any interaction effects were found. To investigate further the two main effects, we run univariate pairwise comparisons using the Holm-Bonferroni adjustment for family-wise error[4].

| | |
|---|---|
| ***SENSI*** | ALL > G+P+HD > P+HD > P > HD+MV > G > ME > HD |
| ***SPECIF*** | ALL > G+P+HD > P+HD > P > HD+ME > G > ME > HD |

*Table 2: Effect of feature set on model performance, with G=Gaze, P=Pupil, HD=Head Distance, ME=Mouse Event.*

**Effects of Feature set.** Table 2 summarizes the results of the pairwise comparisons by ordering feature sets according to their mean sensitivity and specificity over data windows and SMOTE configurations. Bold underlining indicates models for which there are no statistically significant differences. For example, for sensitivity, these is no difference between G+P+HD and HD+P, they are both better than P but worse than ALL.

Results in Table 2 show that combining various data sources works usually better than using a single data source, for both sensitivity and specificity. In particular, the best feature set for both performance measures includes all the features (ALL). This indicates that eye tracking features work well together to predict confusion, and that adding mouse events can lead to significantly better accuracy. It is interesting to note that G+P+HD (the second best model) is not significantly better than P+HD, meaning that adding Gaze data to Pupil and Head Distance features sets does not lead to a significant improvement.

**Effects of SMOTE configuration.** The pairwise comparisons show that:

- Sensitivity is better with SMOTE-200% than with SMOTE-500%, (*t*(198)=3.19, *p* = .002, $\eta^2_p$=.15).
- Specificity is better with SMOTE-500% than with SMOTE-200%, (*t*(198)=2.41, *p* = .017, $\eta^2_p$=.04).

These results show that there is a trade-off between SMOTE-200% and SMOTE-500% in terms of sensitivity and specificity. In particular, generating more synthetic data with SMOTE has a substantial negative effect on the sensitivity (see medium effect size $\eta^2_p$=.15). This is likely due to the fact that synthetic data start to dilute the information captured in the real confusion data points. This effect is inverted for specificity, although the effect size is small ($\eta^2_p$=.04), indicating limited implications in practice.

## 5.1 Model performance

In this section, we analyze in more detail the performance (in terms of sensitivity and specificity) of the classifiers using the ALL and P+HD feature sets and trained over "Full Window"[5] data. We focus on these two feature sets because ALL showed the best overall performance (Table 2) and P+HD is the second best model along with G+P+HD, but it has the advantage to be fully independent from the layout of the visualization. For each of these feature sets and Full window, we report results for both SMOTE-200% and SMOTE-500% (Table 3), since Section 5.1 showed that there is a trade-off between these two configurations.

Overall, results in Table 3 indicate that SMOTE models have very similar specificities, with variations of only about .02 for both ALL and P+HD, whereas SMOTE-200 achieved by far the highest sensitivity[6], reaching .61 for ALL. Thus SMOTE-200 used with all features together appears to be the most promising model to predict user confusion during interaction with ValueCharts. However, layout independent information captured only by pupil and head distance features appears to successfully predict 57% of confusion trials on unseen data, a very encouraging result in terms of building visualization-independent classifiers.

| | | **SMOTE 200%** | **SMOTE 500%** | **Baseline** |
|---|---|---|---|---|
| ***SENSI*** | *ALL* | .61 | .54 | 0 |
| | *P+HD* | .57 | .55 | 0 |
| ***SPECIF*** | *ALL* | .926 | .942 | 1 |
| | *P+HD* | .905 | .921 | 1 |

*Table 3: Performance of ALL and P+HD with Full window.*

## 5.2 Feature importance

Our results show that multiple data sources together (i.e., *ALL Features*) can better predict confusion, than any single source. To ascertain which features best predict confusion, we use the method described in [Liaw and Wiener 2002] to measure feature importance in Random forest classifiers. Table 4 shows the top 10 selected features for our best performing classifier (using the ALL feature set and SMOTE -200%), with relative importance normalized between 0 and 100. In the table, a positive (negative) direction of the effect column (D) indicates that the value of the feature is higher (lower) in confusion than in no-confusion trials.

**Pupil features.** Table 4 shows that overall pupil size features are the most important predictors of confusion, as the top three best features belong to this set. Generally, increase in pupil size is correlated with higher cognitive load [Granholm and Steinhauer 2004]. Consistently with this result, *end*, *max,* and *mean pupil size* are higher in confusion trials, where confused users might be experiencing a higher cognitive load. Also, users in the confusion trails have higher *std.dev pupil size*, which might be an indication of a higher variability in their cognitive load compared to more stable non-confused users. The prominence of pupil-based features as predictors of confusion is especially promising for the generalization of

[4] Actual values for average sensitivity and specificity will be discussed later for specific models.

[5] Windows length had no effect on performance, so either length could be used here.

[6] When differences in specificity is higher, measures such as the practical utility could be used to investigate the trade-off between sensitivity and specificity [Gena 2005].

our work, since these features are independent from the layout of the visualization, and thus may be used to predict confusion in other InfoVis.

**Head distance.** There is one head distance feature among the top 10, namely *std.dev* head distance. Confused users have a higher value for this feature, suggesting that confusion generates more fluctuations in the user's position. Distance to the screen has been shown to be correlated with user engagement [D'Mello and Graesser 2007], thus more fluctuations in the position of confused users might indicate that these users get closer to the screen to better attend to the unclear information before eventually disengaging. As with pupil size, head distance to the screen does not depend on the visualization layout and thus may be a good predictor of confusion using other InfoVis. Also as noted before, head distance may be inferred with a cheaper webcam.

**Gaze features.** Five of the 10 most important features in Table 4 are Gaze features related to attention to the AOI that includes the name of the attributes (*Labels_attr*) in the decision problem (see Fig. 2). These features are *time to last fixation* in the AOI, *longest fixation*, *proportion of time spent*, as well as *transitions* from and to the AOI that encloses the name of the problem alternatives (*Labels alter*). Confused users have higher values for all these features, suggesting that they need to process the names of the attributes more extensively. One possible explanation for this trend is that labels are a source of confusion (e.g., due to ambiguous wording). Another is that looking at names of attributes and alternatives is a way to reduce confusion as text in visualization is meant to support data comprehension. It should be noticed that all Gaze features here relate to a particular AOI, thus are layout dependent. This indicates that although many independent features (e.g., pupil and head distance) are important predictors of confusion, additional information about specific components of the visualizations can improve prediction.

| Features | Set | Score | D |
|---|---|---|---|
| End pupil size | Pupil | 100 | + |
| Max pupil size | Pupil | 88 | + |
| Stddev pupil size | Pupil | 67 | + |
| Labels_attr: proportional time spent | Gaze | 45 | + |
| Stddev distance | Head | 39 | + |
| Mean pupil size | Pupil | 37 | + |
| Labels_vis: time to last fixation | Gaze | 36 | + |
| Labels_attr: time to last fixation | Gaze | 33 | + |
| Labels_attr: num of transitions to Labels_alter | Gaze | 29 | + |
| Labels_alter: number of transitions to Labels_attr | Gaze | 26 | + |

*Table 4: Top 10 features for predicting confusion·*

## Conclusion

If confusion could be predicted and resolved in real time, user experience and satisfaction with InfoVis would be greatly improved. In this paper, we focused on predicting occurrences of confusion during the interaction with ValueChart, an interactive visualization to support multi-criteria preferential choice. To this end, we leveraged a user study that collected ground truth labels for confusion, along with eye tacking and interaction data. Then, we compared various combinations of these data sources to train Random forest classifiers for confusion, and technically had to deal with data imbalance.

Our results show that eye tracking is valuable to predict confusion in real time. Remarkably, we found that 61% of the occurrences of confusion can be predicted, while getting a false positive rate of only 7.4%. More tellingly, when we examine the most important features used by the classifiers, it appears that our models are able to capture aspects of the interaction that are very plausibly related to confusion. Furthermore, some of these features may well generalize to other visualizations. For instance, we found that features of pupil size (which are independent of the layout of the current visualization) are strong predictors of confusion, consistently with the fact that pupil size is correlated to cognitive load, which plausibly correlates with confusion. Another strong visualization-independent predictor was a feature related to head distance. Again this makes sense, because confusion can affect engagement with a task, which has been shown to be predictable by head distance. Additionally, prominent gaze features reveal differences in user's attention to the labels of the visualization among confused and non-confused users; however, these features may not generalize so easily to other visualizations.

To increase the performance of our models, future work includes improving our features and models selection via ensemble modeling, leveraging additional features such as facial expressions, and using past observed data to optimize prediction for each individual user. Another thread of future work relates to further investigating the generalizability of our findings to other visualizations, as well as researching how confusion can be addressed once predicted.

## Acknowledgement

This research was funded by NSERC Grant 227795-10. We thank Enamul Hoque and Dereck Toker for their help with data collection, and Dereck for his help in revising the paper.

## References

[Baker *et al.*, 2012] R. Baker, S.M. Gowda, M. Wixon, *et al.* Towards Sensor-Free Affect Prediction in Cognitive Tutor Algebra. In *Proc. of EDM 2012*, 126–133, 2012.

[Bixler and D'Mello, 2015] R. Bixler and S. D'Mello. Automatic Gaze-Based Prediction of Mind Wandering with Metacognitive Awareness. In *Proc. of UMAP 2015*, 31–43, 2015. Springer.

[Bosch *et al.*, 2015] N. Bosch, S. D'Mello, R. Baker, *et al.* Automatic Prediction of Learning-Centered Affective States in the Wild. In *Proc. of IUI 2015*, 379–388. ACM.

[Chawla *et al.*, 2002] N.V. Chawla, K.W. Bowyer, L.O. Hall, *et al.* SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.

[Conati *et al.*, 2014] C. Conati, G. Careninni, B. Steichen, *et al.* Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. *Computer Graphics Forum*, 33(3):371–380, 2014.

[Conati *et al.*, 2013] C. Conati, E. Hoque, D. Toker, *et al.* When to Adapt: Predicting User's Confusion During Visualization Processing. In *Proc. of WAUV*, 2013.

[D'Mello and Graesser, 2007] S. D'Mello and A. Graesser. Mind and Body: Dialogue and Posture for Affect Prediction in Learning Environments. In *Proc. of AIED 2007*, 161–168, 2007.

[Fawcett, 2006] T. Fawcett. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8):861–874, 2006.

[Gena, 2005] C. Gena. Methods and techniques for the evaluation of user-adaptive systems. *Knowl. Eng. Rev.*, 20(1):1–37, 2005.

[Granholm and Steinhauer, 2004] E. Granholm and S.R. Steinhauer. Pupillometric measures of cognitive and emotional processes. *Int. J. Psychophysiol.*, 52(1):1–6, 2004.

[Holmqvist and Nyström, 2015] K. Holmqvist and M. Nyström. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. OUP, 2015.

[Huang *et al.*, 2015] D. Huang, M. Tory, B.A. Aseniero, *et al.* Personal Visualization and Personal Visual Analytics. *IEEE Trans. Vis. Comput. Graph.*, 21(3):420–433, 2015.

[Iqbal *et al.*, 2005] S.T. Iqbal, P.T. Adamczyk, X.S. Zheng, *et al.* Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proc. of ACM SIGCHI 2005*, 311–320, 2005.

[Jang *et al.*, 2014] Y.-M. Jang, R. Mallipeddi, and M. Lee. Identification of human implicit visual search intention based on eye movement and pupillary analysis. *User Model. User-Adapt. Interact.*, 24(4):315–344, 2014.

[Jaques *et al.*, 2014] N. Jaques, C. Conati, J.M. Harley, *et al.* Predicting Affect from Gaze Data during Interaction with an Intelligent Tutoring System. In *Proc. of ITS 2014*, 29–38, 2014. Springer.

[Kotsiantis *et al.*, 2006] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.*, 30(1):25–36, 2006.

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

[Lallé *et al.*, 2015] S. Lallé, D. Toker, C. Conati, and G. Carenini. Prediction of Users' Learning Curves for Adaptation While Using an Information Visualization. In *Proc. of IUI 2015*, 357–368, 2015. ACM.

[Lee *et al.*, 2016] S. Lee, S. Kim, Y. Hung, *et al.* How do People Make Sense of Unfamiliar Visualizations? A Grounded Model of Novice's Information Visualization Sensemaking. *IEEE Trans. On Vis. Comput. Graph.*, 22(1):499–508, 2016.

[Liaw and Wiener, 2002] A. Liaw and M. Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[Lim *et al.*, 2015] Y. Lim, A. Ayesh, and M. Stacey. Using Mouse and Keyboard Dynamics to Predict Cognitive Stress During Mental Arithmetic. *Intelligent Systems in Science and Information*, 335–350, 2015.

[Liu *et al.*, 2010] C. Liu, R.W. White, and S. Dumais. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proc. of ACM SIGIR 2010*, 379–386, 2010.

[Muldner et al, 2010] K. Muldner, W. Burleson, and K. VanLehn. "Yes!": using tutor and sensor data to predict moments of delight during instructional activities. In *Proc. of UMAP 2010*, 159–170, 2010.

[Nadkarni and Gupta, 2007] S. Nadkarni and R. Gupta. 2007. A task-based model of perceived website complexity. *Mis Q.*, 31(3):501–524.

[Nazemi *et al.*, 2014] K. Nazemi, W. Retz, J. Kohlhammer, *et al.* User similarity and deviation analysis for adaptive visualizations. In *Human Interface and the Management of Information*, 64–75, 2014. Springer.

[Pentel, 2015] A. Pentel. Patterns of Confusion: Using Mouse Logs to Predict User's Emotional State. In *Proc. of the PALE Workshop*, 40–45, 2015.

[Porayska-Pomsta *et al.*, 2013] K. Porayska-Pomsta, M. Mavrikis, S. D'Mello, *et al.* Knowledge elicitation methods for affect modelling in education. *Int. J. Artif. Intell. Educ.*, 22(3):107–140, 2013.

[Steichen *et al.*, 2014] B. Steichen, C. Conati, and G. Carenini. Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *ACM Trans. Interact. Intell. Syst.*, 4(2):11, 2014.

[Wongsuphasawat *et al.*, 2012] K. Wongsuphasawat, C. Plaisant, *et al.* Querying event sequences by exact match or similarity search: Design and empirical evaluation. *Interact. Comput.*, 24(2):55–68, 2012.

[Wu, 2015] M. Wu. *Inferring user cognitive abilities from eye-tracking data*. University of British Columbia, MsC Thesis, 2015.

[Yelizarov and Gamayunov, 2014] A. Yelizarov and D. Gamayunov. Adaptive Visualization Interface That Manages User's Cognitive Load Based on Interaction Characteristics. In *Proc. of ACM VINCI 2014*, 1–8, 2014.

[Yi, 2008] J.S. Yi. *Visualized decision making: development and application of information visualization techniques to improve decision quality of nursing home choice*, Georgia Institute of Technology. PhD Thesis, 2008.