Fast and Comprehensive Extension to Intention Prediction from Gaze

Hana Vrzakova
University of Eastern Finland
School of Computing
hana.vrzakova@uef.fi

Roman Bednarik
University of Eastern Finland
School of Computing
roman.bednarik@uef.fi

ABSTRACT

Every interaction starts with an intention to interact. The capability to predict user intentions is a primary challenge in building smart intelligent interaction. We push the boundaries of state-of-the-art of inferential intention prediction from eye-movement data. We simplified the model training procedure and experimentally showed that removing the post-event fixation does not significantly affect the classification performance. Our extended method both decreases the response time and computational load.

Using the proposed method, we compared full feature sets to reduced sets, and we validated the new approach on a complementary set from another interaction modality. Future intelligent interfaces will benefit from faster online feedback and decreased computational demands.

Author Keywords

Intentions; Prediction; Eye-tracking; SVM; Gaze-augmented interaction; Dwell-time

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Prediction of user interests and intention to interact is the primary task of user interface designers. Best UI designs are those that tap into users' preferences and provide a seamless interaction where the interface *'knows'* what are the intentions of the user at any time. While anticipating future interactions, designers can impersonate a typical user, can try to estimate his model in head, gain understanding of the needs, and express that in terms of the design of the interface that matches the interaction model of the user. If they succeed, the interface is perceived as natural, user friendly, responsive, immersive and intuitive, to name few.

An everyday experience unfortunately indicates that such user interfaces are rare. One reason for it is that the designers fail to engineer the proper interaction model and because of the mismatches between the current user perception of the system and the real state of the system. If a user interface can effectively predict that a user wants to interact in a certain way even though the current state of the system does not expect such to happen, interaction errors can be avoided. For instance, misplaced interactions, such as expecting to type in an input box but not having the cursor at the input box, can be efficiently corrected when an intention to type can be predicted early enough.

All interactive actions begin with an intention to interact. Specifically, the formation of the intention to explicitly interact is a stage preliminary to interaction [13]. For example, to press a button a user has to first formulate an intention to interact and then execute the hand movement and finger flex to issue the button press. In this paper we deal with the deep detailed level of interaction in terms of predicting the intentions to interaction.

Eye tracking as source for user modeling

Eye-tracking data can be used to discover user's cognitive state [8, 14], workload [2, 1], expertise [9] or to predict the context of interaction [11, 6]. Eye-tracking is also expected to become a ubiquitous interaction technique [12, 5]. If eye-tracking is indeed going to be a standard source of user data, the implicit behavioral information can be used for modeling of user states.

Previous work on intention prediction has shown that employing eye-movements and machine learning is a feasible modeling technique to achieve good levels of prediction in human-computer interaction. Bednarik et al. formulated a machine learning pipeline for eye-tracking data that performs training of a classifier to detect whether a user, engaged in gaze-augmented interactive problem solving, aims to change the state of the problem using a button press [4]. Using their framework, they achieved a classification accuracy of 76% (AUC = 0.81). Thus, intention prediction is achievable with levels far above the level of chance, although the total training time reported was over 180 hours.

In this paper, we report on a systematic study to advance the state-of-the-art of automatic inferential intention prediction for HCI. We 1) employ another modality to evaluate the generalizability of the pipeline, 2) present effects of simplification of the training, 3) investigate new options of feature extraction, and 4) compare the performance of the feature sets of the state-of-the-art system with performance based on reduced feature sets.

IUI 2013 Workshop: Interacting with Smart Objects, March 18, 2013, Santa Monica,CA, USA

In particular, we compare performance differences in intention prediction for gaze-augmented and traditional mouse-based interaction. With an objective to significantly reduce training time, we designed a less comprehensive training and evaluate the effects of the simplification on the performance.

The original study has employed *fixational sequences* centered around the observed intention. It implies that in real-time implementations the prediction component would be able to report on a predicted intention with some delay. The underlying question, however, concerns maximalizing the chance of discovering intentions from short fixational sequences. Optimally, we would wish to be able to predict an incoming interaction *before* it happens. Therefore, we systematically shift the extracted sequences before and after the interactive action and compare the effects on the prediction performance.

Finally, we perform a standard feature selection to reduce available feature space, by analyzing inter-feature correlations.

METHOD

Training dataset: Interactive problem solving

The datasets that we employ in this research were originally collected for another purposes and has been described in [3]; the original study examined the effects of interaction modality on problem-solving activities.

Figure 3 presents a studied user interface from the original study. The task of the user was to arrange a shuffled set of tiles into a required order. As users engaged in the problem solving through various interaction modalities, the original studies have discovered that gaze-augmented interaction is superior over the traditional mouse-based interaction.

The data were collected in a quiet usability laboratory. The eye movements of the participants were collected using a Tobii ET1750 eye-tracker, sampling at 50Hz. Each participant interacted with the interface individually and participants were required to think aloud. There were altogether three sessions from which data has been collected.

Here we use two datasets from those interactions: first, the same gaze-augmented interaction dataset as employed in the benchmark report by Bednarik et al. [4]. Second, the difference here is that the evaluation of the new method employs also a dataset containing mouse-controlled interactions. Thus, in the first dataset gaze is used in a bidirectional way, while in the mouse-based problem-solving gaze is used only for perception of the problem-space.

The button press in both conditions sets the boundaries for the fixational sequence extraction. The corresponding sequence of eye tracking data is related to this event. The sizes of the extracted datasets are shown in Table 1.

Extension of prediction method

The experiments in this study take as a baseline the prediction framework from [4]. The prediction framework performs detection of intentions using fixational sequences wrapped around the interaction event. It employs numerous eye-tracking features computed from the sequences, and cross-validation for prediction model training. A parameter grid search is employed and Support Vector Machine is used as a classifier.

Table 1. Dataset distributions according to interaction style

Type of interaction	Intent [n]	Non-Intent [n]	Total [n]
Gaze Augmented	2497	22119	24616
	10.14%	89.86%	100%
Mouse only	2823	18714	21537
	13.11%	86.89%	100%

In this work, we propose two extensions to the previous work. Figure 1 illustrates the prediction pipeline and introduces the main contributions of the present work as presented in the following sections.

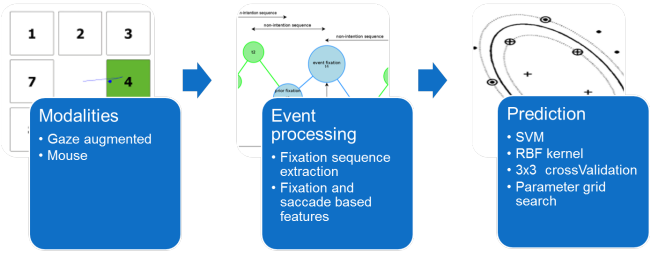

Figure 1. Extensions in prediction pipeline.

Disregarding future information

The first modification concerns the extraction of the fixational sequences. The original work focused on wrapping the feature extraction around the so-called *event* fixation: whole sequence consisted of the event fixation reflecting the interaction event, one fixation before and one fixation after the event.

Here we introduce *pre-event* and *post-event* fixations. Using this scheme, illustrated in Figure 2, we created three datasets: one consisting of sequences composed from two pre-event and one event fixations (denoted hereafter by '2+1+0'), one consisting of pre-event, event, and post-event fixations (1+1+1), and one of one event and two post-event fixations (0+1+2). Such settings, we believe, may reveal contribution of fixations surrounding interaction events.

The second expansion focuses on the type of computed features. We employ fixation and saccade features only and disregard pupil-dilation based features. Although prior research proved a link between pupil dilation and cognitive processes [1], it has also revealed a tangible time delay between cognition and pupillary response [7]. Such delay would deteriorate

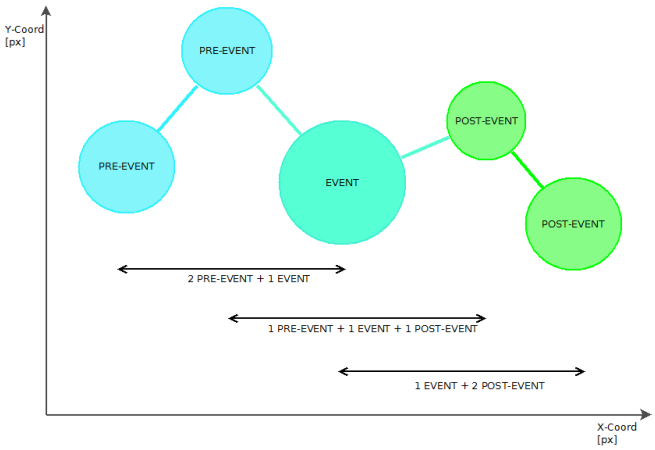

Figure 2. Fixational sequences: Pre-event, event and post-event fixations

the performance of an eventual real-time classifier. The fixation and saccade based features are presented in Tables 2 and 3.

Table 2. Eye movements features computed from fixations. Adopted from[4]

Eye movement feature	Description
Mean fixation duration	The average time of fixation duration in the observed sequence
Total fixation duration	Sum of fixation durations in the observed sequence
Event fixation duration	Duration of the fixation for the ongoing interaction
Prior fixation duration	Duration of the fixation before intention occurrence

Table 3. Eye movements features computed from saccades. Adopted from [4]

Eye movement feature	Description
Mean saccade duration	The average saccade duration in the observed sequence
Total saccade duration	Sum of saccade durations in the observed sequence
Last saccade duration	Duration of the fixation before event occurrence
Mean saccade length	The average distance of saccade in the observed sequence
Total saccade length	Sum of saccade distances in the observed sequence
Last saccade length	Distance of the saccade before event occurrence
Mean saccade velocity	The average speed of saccades in the observed sequence
Last saccade velocity	Speed of the saccade before event occurrence
Mean saccade acceleration	Acceleration of saccade during the observed sequence

Faster training

Third, *simplified* the parameter search in prediction model training by reducing the number of folds in the two nested cross validations from 6 x 6 to 3 x 3. Such settings reduce computational time. More importantly we investigate whether it affects the classifier performance.

Fourth, we created an additional dataset by filtering out correlated features. Such reduced dataset may have comparable or better performance under lower computational costs.

Baseline settings

For comparison purposes, we created a *balanced* dataset of intent and non-intent feature vectors. We used all the intentions and randomly chose a corresponding number of non-intention feature vectors (see Table 1). In real interaction, the proportion of intentions is much lower, however, balanced experimental settings serve for baseline comparison and show the limitations of the weight settings in case of an unbalanced training dataset.

The remaining settings (parameter grid search and SVM kernel) were kept the same as in the prior study [4].

RESULTS

The systematic evaluation, reported here, presents 18 experiments, with a total duration over 135 hours of computational time, which presents reduction around 30% compared to prior study in [4], when using a comparable hardware.

A typical processed sequence of fixations containing an event is shown in blue color in Figure 3.

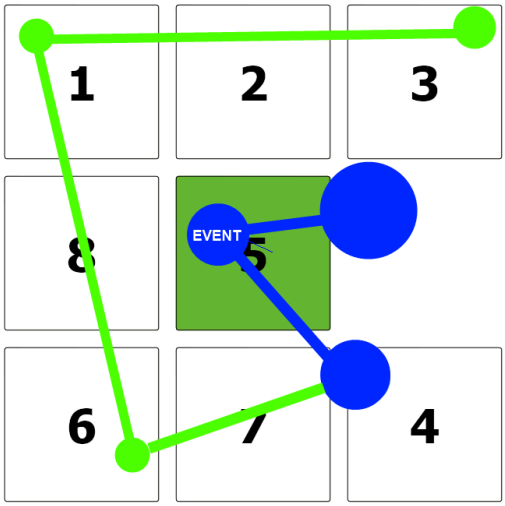

Figure 3. Typical intent (blue) and non-intent (green) fixation sequences. The relationship between features computed from the sequences are shown in Figures 4 and 5

Understanding how much each feature contributes to the class recognition belongs to the features selection problems and lead to another computationally demanding tasks. For estimation of such contribution, Figure 4 and Figure 5 demonstrate a percentage change of averaged features, when compared to the non-intention ones, and how observed interaction style influenced the ratio. A baseline indicates that intention and non-intention feature vectors would be same, positive ratio shows greater mean values of the intention-related features while negative presents the smaller ones.

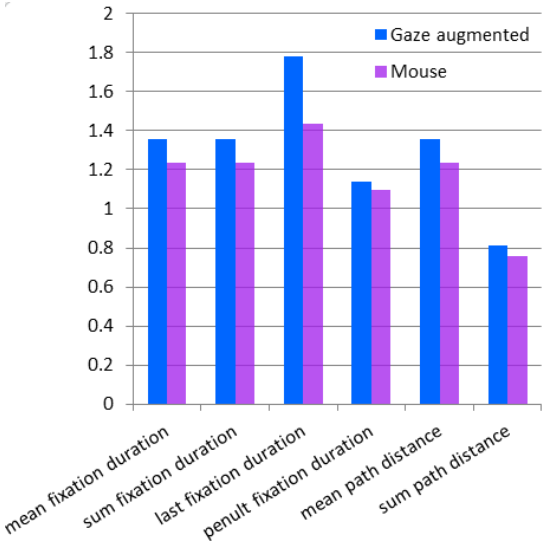

Figure 4. Effects of intentions on fixation based features. During intentional interaction, fixation derived metrics increased compared to baseline (non-intentional interaction). The comparison of interaction styles introduced higher increase in the gaze-augmented interface.

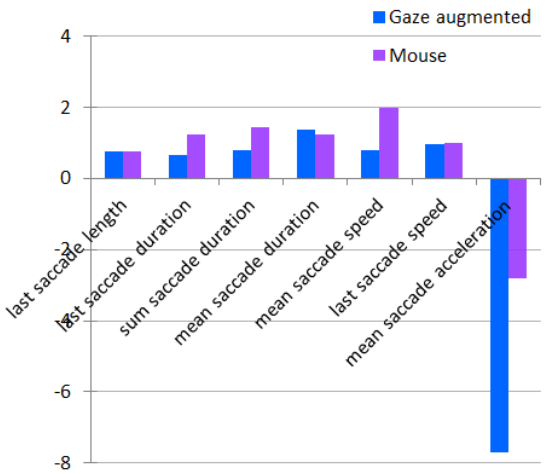

Figure 5. Influence of intentions on saccade based features. Intentions to interact increased metrics of saccade duration and speed and decreased saccade acceleration, when compared to baseline (non-intentions). The mouse interaction style reflects more in the duration and speed features, while acceleration corresponds with the gaze-augmented interaction.

Table 4 shows an overview of all experiments. We report on Area Under the Curve (AUC) as the primary performance measure. In case of the balanced data, also accuracy is a reliable metric of classification performance.

The performance of the classifiers is comparable to the baseline results achieved previously in [4]; best performance on an imbalanced data was AUC of 0.79 which is just 0.02 below the 0.81 reported before. However, the best performance here was achieved using much simpler training procedure. The effect of feature selection was minor, however noticeable.

Effects of Fixation sequence

A comparison of the processing pre-event and post-event fixational sequences showed minor differences in each training group. In gaze-augmented The highest performance was reached up to AUC of 0.8 in gaze augmented interaction (SVM.C = 94.868, SVM.Gamma = 3.684E-7), and AUC of 0.73 in the traditional mouse interface (SVM.C = 30.800, SVM.Gamma = 1.0E-8). Although in several cases of mouse modality AUC resulted better performance for post-event dataset, the better performance in gaze augmented interface was gained using pre-event (2+1+0 and 1+1+1) fixational sequences.

Predictability of intentions across modalities

A comparison of interaction modality showed that intention prediction was better performed using gaze-augmented interaction rather than mouse-based one. The best performance for gaze-augmented interface reached up to AUC 0.8, while the best mouse-based prediction was still 0.07 lower. Such results indicate that interaction intention prediction is tightly dependent on the observed modality, and prediction model needs training for each modality separately.

DISCUSSION

This paper presented two contributions. The main novelty presented here is in showing that interaction intention prediction from gaze does not need to rely on post-event fixations. This finding has important implications, both on the research of understanding of interactions from physiological signals and on the applications and implementations of the inference algorithms in real time.

We reported the cross-validation results of the extended intention prediction pipeline. Here we compared them to the prior baseline study, reported in [4].

Predicting interaction before actions

The findings show that it is not necessary to postpone action detections until post-event data becomes available. This can be considered as a breakthrough result, give the fact that research that employs EEG signals for detection of interaction errors reports the shortest achievable time for a computing system to realize action to be about 150 - 300ms after the user-event [10].

A question arises regarding the information contained in the fixational sequences. Where one should look for a reliable source of information about interaction intention? According to Figures 4 and 5 and the ratios of averaged feature vectors, the two interaction modalities resulted in observable differences between averaged intention and non-intention feature vectors. In other words, the gaze behavior around interactive actions differed across modalities. We observed that gaze-augmented interface affected more the features related to fixation, whereas the interface with the mouse influenced saccade based features. Therefore, the answer to the question seem to depend on the interaction modality in use.

Table 4. Overview of results

Modality	Training	Fixation sequence	AUC	Accuracy	Recall	Precision
Gaze augmented	State of the art. Adapted from [4]	1 + 1 + 1	0.81	0.76	0.69	0.31
Gaze augmented	Simplified	2 + 1 + 0	0.78	0.82	0.54	0.29
		1 + 1 + 1	0.78	0.80	0.57	0.27
		0 + 1 + 2	0.79	0.82	0.57	0.29
	Simplified + Without correlated features	2 + 1 + 0	0.72	0.75	0.53	0.21
		1 + 1 + 1	0.72	0.75	0.54	0.21
		0 + 1 + 2	0.75	0.77	0.54	0.23
	Simplified + Balanced	2 + 1 + 0	0.77	0.71	0.66	0.74
		1 + 1 + 1	0.80	0.73	0.72	0.73
		0 + 1 + 2	0.78	0.72	0.67	0.75
Mouse	Simplified	2 + 1 + 0	0.69	0.70	0.65	0.23
		1 + 1 + 1	0.72	0.68	0.65	0.23
		0 + 1 + 2	0.72	0.67	0.65	0.23
	Simplified + Without correlated features	2 + 1 + 0	0.67	0.58	0.74	0.19
		1 + 1 + 1	0.69	0.62	0.66	0.20
		0 + 1 + 2	0.71	0.64	0.70	0.22
	Simplified + Balanced	2 + 1 + 0	0.70	0.64	0.55	0.67
		1 + 1 + 1	0.71	0.66	0.65	0.66
		0 + 1 + 2	0.73	0.66	0.59	0.69

Reduction of the computational load

The second contribution of this study lies in showing that reducing the comprehensive search for optimal parameters during training is justified by minimal decrease in classification performance. This is a major improvement, because the decreased complexity and costs of the training lead to less computational load. In sum, a less comprehensive search in the feature space does not necessarily imply worse performance of the classification when using SVM classifiers.

Considering the implications of the findings on the real-time implementations, we have virtually removed the need to delay classification decisions till post-even fixations. Thus, an effective inference can be carried out at nearly the real-time of the event. We are currently investigating possibilities to even a further shift – in the sense of employing more data from the past– however, there are new challenges arising. For example, the previous events that are close to the event of interest create overlapping data and thus ground truth labeling is difficult.

Finally, to demonstrate the robustness and generalizability of the new approach, we evaluated its performance on a complementary dataset. Although the features differ because of a different interaction modality, the performance of the intention classification pipeline only decreases by about 5-9% on AUC.

Applications of intention inference

The research presents eye-tracking as a feasible and fast method for intention classification. Although the datasets on which we developed the methods have been captured from a rather traditional WIMP paradigm, we believe that in contexts beyond a computer desktop our approach can as well be applied.

Event though the current wearable eye-trackers do not achieve high sampling rates, and thus the temporal resolution is low to allow accurate identification of fast eye-movements, future technologies will likely be able to overcome this drawback. Then, methods such as ours can be used for detection of user intention to interact with surrounding objects. For a pervasive interaction, not only the objects can be made gaze-aware [15], but can be made even mode intelligent by sensing the nuances of user interaction with them.

CONCLUSION AND FUTURE WORK

The ability to predict user intentions is one of the primary challenges in building smart intelligent interfaces. Our study extends the argument that eye movements can reveal interaction intentions and their relationship to interaction style using intention prediction model.

In comparison to prior research, we lowered computational demands of 30% using balancing dataset, reducing number of folds in cross validations, and removing correlated features. Even though a comparison of classification performance revealed a decreased ability to differentiate between intentions and non-intentions, such approach motivates for further research since the overall classification performance was reduced just in acceptable units of AUC. For future real-time classifications, methods of optimized prediction are more promising than the demanding parameter search in a large feature space.

REFERENCES

1. Bailey, B. P., and Iqbal, S. T. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Trans. Comput.-Hum. Interact. 14*, 4 (Jan. 2008), 21:1–21:28.

2. Bartels, M., and Marshall, S. P. Measuring cognitive workload across different eye tracking hardware platforms. In *Proceedings of the Symposium on Eye*

Tracking Research and Applications, ETRA '12, ACM (New York, NY, USA, 2012), 161–164.

3. Bednarik, R., Gowases, T., and Tukiainen, M. Gaze interaction enhances problem solving: Effects of dwell-time based, gaze-augmented, and mouse interaction on problem-solving strategies and user experience. *Journal of Eye Movement Research 3*, 1 (2009), 1–10.
4. Bednarik, R., Vrzakova, H., and Hradis, M. What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, ACM (New York, NY, USA, 2012), 83–90.
5. Bulling, A., and Gellersen, H. Toward mobile eye-based human-computer interaction. *Pervasive Computing, IEEE 9*, 4 (2010), 8–12.
6. Bulling, A., Roggen, D., and Troster, G. What's in the eyes for context-awareness? *Pervasive Computing, IEEE 10*, 2 (2011), 48–57.
7. Einhäuser, W., Koch, C., and Carter, O. L. Pupil dilation betrays the timing of decisions. *Frontiers in human neuroscience 4* (2010).
8. Eivazi, S., and Bednarik, R. Inferring problem solving strategies using eye-tracking: system description and evaluation. In *Proceedings of the 10th Koli Calling International Conference on Computing Education Research*, Koli Calling '10, ACM (New York, NY, USA, 2010), 55–61.
9. Eivazi, S., Bednarik, R., Tukiainen, M., von und zu Fraunberg, M., Leinonen, V., and Jääskeläinen, J. E. Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, ACM (New York, NY, USA, 2012), 377–380.
10. Ferrez, P. W., and Millán, J. D. R. You are wrong!: automatic detection of interaction errors from brain waves. In *Proceedings of the 19th international joint conference on Artificial intelligence*, IJCAI'05, Morgan Kaufmann Publishers Inc. (San Francisco, CA, USA, 2005), 1413–1418.
11. Hradis, M., Eivazi, S., and Bednarik, R. Voice activity detection from gaze in video mediated communication. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, ACM (New York, NY, USA, 2012), 329–332.
12. Jacob, R. J. K., and Karn, K. S. Commentary on section 4. eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*. Elsevier Science, 2003, 573–605.
13. Norman, D. A. *The Design of Everyday Things*. Basic Books, New York, 2002.
14. Simola, J., Salojärvi, J., and Kojo, I. Using hidden markov model to uncover processing states from eye movements in information search tasks. *Cogn. Syst. Res. 9*, 4 (Oct. 2008), 237–251.
15. Vertegaal, R., and Shell, J. Attentive user interfaces: the surveillance and sousveillance of gaze-aware objects. *Social Science Information 47*, 3 (2008), 275–298.