

Quiet Eye Affects Action Detection from Gaze more than Context Length

Hana Vrzakova and Roman Bednarik

University of Eastern Finland,
PO 111, 80101 Finland
{hanav, roman.bednarik}@uef.com

Abstract. Every purposive interactive action begins with an intention to interact. In the domain of intelligent adaptive systems, behavioral signals linked to the actions are of great importance, and even though humans are good in such predictions, interactive systems are still falling behind. We explored mouse interaction and related eye-movement data from interactive problem solving situations and isolated sequences with high probability of interactive action. To establish whether one can predict the interactive action from gaze, we 1) analyzed gaze data using sliding fixation sequences of increasing length and 2) considered sequences several fixations prior to the action, either containing the last fixation before action (i.e. the quiet eye fixation) or not. Each fixation sequence was characterized by 54 gaze features and evaluated by an SVM-RBF classifier. The results of the systematic evaluation revealed importance of the quiet eye fixation and statistical differences of quiet eye fixation compared to other fixations prior to the action.

Keywords: Action, Intentions, Prediction, Eye-tracking, SVM, Mouse interaction, Problem solving

1 Introduction

Understanding users, their interests and actions computationally is key for tapping into user's needs and provision of seamless interactions where the interactive systems *knows user's intentions*. A good interface design gives an impression of being able of anticipating future interactions because designers succeeded in understanding of the model in head and the user's needs. If they succeeded, the interface is perceived as natural, user friendly, responsive, immersive and intuitive.

Everyday experience unfortunately indicates that such user interfaces are still scarce. Current interactive systems typically do not contain mechanisms for prediction of the user's actions, and consequently restrict them from implementing their intentions effectively. It is however not only the designers who have to anticipate the user's actions; the interface itself needs to play an active role and just in time, or even *ahead of time*, adapt to the changing user needs proactively.

There are numerous benefits if an interface is able to *predict* that the user wants to interact with it, but one of the primary motivations is to mitigate consequences of interaction errors. For example, when a system assumes the cursor to be located at the respective field for user input – such as when starting to type a search query– a truly proactive interface needs to be able to detect user’s *intention to input*. Detection can then trigger automatic assistance to adjust the cursor location to avoid an unintended error. The eventual errors could be avoided if the intention to type was predicted early enough.

This work considers modeling and automatic detection of actions in human computer interaction. All interactive actions begin with an intention to interact. Specifically, the formation of the intention to explicitly interact is a stage preliminary to interactive activity [21] and part of larger planning activities. For instance, to press a button a user has to first internally formulate an intention to interact, then execute the hand movement toward the button, and finally, flex the finger to issue the button press. Finger flexes, however, are not the most reliable indicators of intention, since they embody the post-intention activity that is merely mechanically executed (a button was pressed after an intention to press it occurred). The novelty of this work is to model computationally the stages preliminary to actions. In this work we consider an intention as an entity for action recognition. To access the plan formation activities, we employ eye-movement analysis as a proxy to cognitive processes related to action formation.

Eye movements have for long been established as a window to human cognition, including planning and motor action [16, 23, 12]. As indicators of voluntary and involuntary attention, eye movements can reveal intention origins and uncover the mechanisms of action planning, execution and evaluation. In this work we examine mouse interaction, we focus on action detection from user’s eye-movements and lay down pathways towards interaction design enhanced by user’s actions.

1.1 Gaze in proactive user interaction

The problem of plan recognition using a computational agent have for long been a central issue in artificial intelligence research [18]. In this work we deal with intention on the level of motor-interactive action. While there are higher- and more sophisticated levels of intentions, such as social intentions (A wants to leave a good impression on B), here we explore intentions at the lower level of action implementation [13]. Our overall goal is to develop a system that would be able to reliably and effectively perform online intention detection.

We employ eye-gaze as a source of intention information since gaze reliably indicates the person’s focus of visual attention and can be unobtrusively tracked. Gaze is also proactive as it reflects the anticipated actions when gathering critical information before performing actions [12], and therefore modeling of proactive gaze can potentially lead to prediction of the resulting actions. Understanding this level of interactive action planning is key for implementation of proactive intelligent systems, in particular, for avoidance of interaction slips [21], Midas Touch effect [15], action slips [13] and human errors in interface design [22].

1.2 Eye tracking in interaction modeling

Eye-tracking data can be used to discover user’s cognitive states [6, 10, 24], workload [1, 2], expertise [19, 11, 5] or to predict the context of interaction [14, 9] and aspects of learning with intelligent tutoring systems [17, 7]. Eye-tracking is also expected to become a ubiquitous interaction technique [15, 8, 27]. If eye-tracking is indeed going to be a pervasive source of user data, the implicit behavioral information can be used for modeling of user states.

Because of the voluminous eye-tracking stream, current research employed customized machine learning techniques as a feasible modeling approach and achieved acceptable levels of predictions. Existing research reached acceptable predictions in human-computer interaction, such as mind-wandering [6] and cognitive abilities [25].

Starting from the pioneering work that adopted a standard classification method for prediction of problem solving states [5], recent research investigated the nuances of eye-tracking pattern-recognition systems in terms of data pre-processing methods [3] or classifier training approaches [28].

In prior work, we and others explored a machine learning pipeline for eye-tracking data that performs training of a classifier to detect various states and individual characteristics of a user. In [5] we presented a machine learning pipeline for intention detection from gaze. Later, we improved the efficiency of the method [28] and evaluated various options for data processing, such as effect of the data before and after an intention occurs, and simplified classifier training.

Although the eye-tracking pattern-recognition systems presented so far achieve classification accuracies far above the chance levels, there are several technical, methodological, and practical questions that motivate the improvements of the prediction pipeline.

In this work we deal with one of the essential questions, namely, how much information an automated modeling system needs to make a reliable decision?

2 Methods

We explore interactive actions during problem solving and recording framework of 8Puzzle[4]. In the 8Puzzle game (Figure 1), users re-arrange the moving tiles into final configuration using a traditional computer mouse. The interaction in detail consisted of moving the mouse cursor onto the tile to be moved, pressing a button, upon which the tile moved to the empty position. The final mouse click represents the interactive action and was recorded with the timestamp in the gaze signal data and represent the ground truth.

2.1 Experimental task and procedure

The task was to arrange the originally shuffled eight tiles into a required target configuration. There were altogether one warm-up trial (data was excluded from the analysis) and three sessions from which data has been collected. On

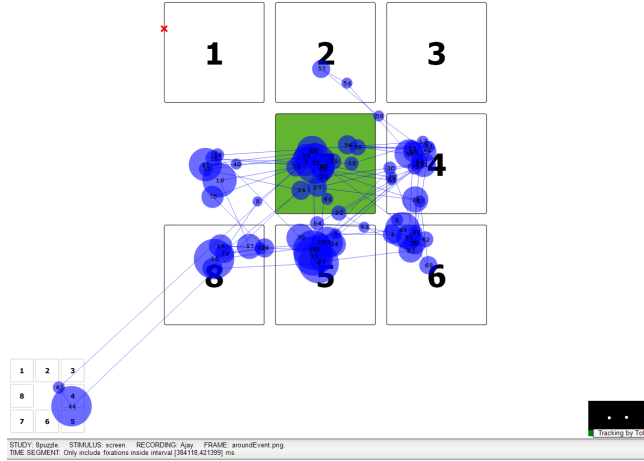


Fig. 1: The interface of 8Puzzle. The shuffled tiles with numbers illustrate a current configuration of the game and the lower left corner the target configuration. Blue circles represent the participant’s fixations in sequence; the visualization of the gaze was hidden to the participants during the experiment.


Users	Modality	Experiment					Actions
		Calibration	Warm-up	Task 1	Task 2	Task 3	
11		Calibration	Warm-up	Task 1	Task 2	Task 3	2 595

Fig. 2: Overview of the experimental design

average a session took about five minutes, and participants were instructed to solve the puzzle till the end. Each participant interacted with the interface individually and participants were motivated to think aloud. Figure 2 summarizes the experimental design, number of participants and the size of the collected dataset.

2.2 Participants and apparatus

The mouse-based experiment consisted of 11 participants (5 male and 6 female) with normal or corrected-to-normal vision, in the 24-47 age range (mean age = 30.36, sd=7.90).

The experiments were conducted in a quiet usability laboratory. Participants’ eye movements were recorded binocularly using a Tobii ET1750 eye-tracker, sampling at 50Hz. The default settings of event identification were set for fixation detection (ClearView, fixation radius 30px, minimal fixation duration 100ms). The

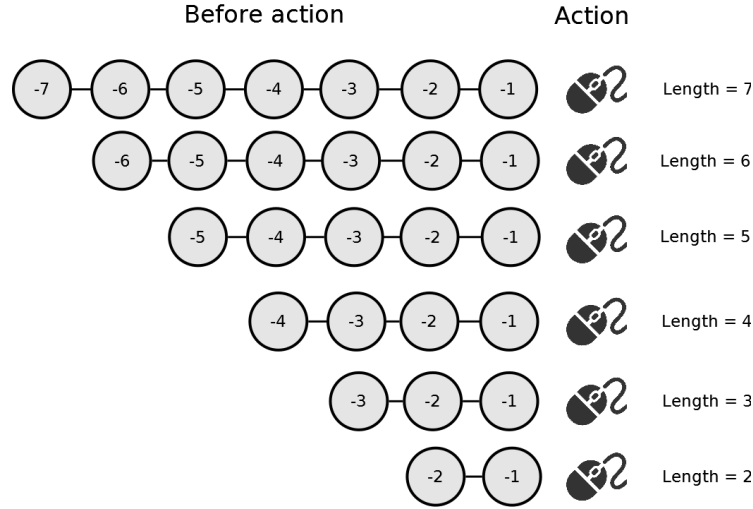


Fig. 3: Scheme of the sequence lengths. The sequence ending right before the mouse click has been annotated as *action*, the rest of sequences before and after the mouse click as *non-action*.

mouse button press was automatically logged into the stream of eye-tracking data and sets the boundaries for action prediction. The following analysis and classification were performed using custom Python scripts, RapidMiner and SPSS.

2.3 Analysis of data

To understand how much data is needed to reliably predict the upcoming action, we employ gaze fixations as a unit of analysis. Fixations are indices of cognitive processing, and extracted short and long sequences of gaze fixations captured before the mouse click should correspond to various stages of cognitive activities. Here, we systematically shifted the fixation sampling window through the data and created following datasets, illustrated in Figure 3. All data was sequenced and annotated as either *action* (action happened after the last fixation in the actual sequence) or *non-action* (other sequences where no interaction occurred). The sequences were extracted with one-fixation overlap to emulate fixation processing as implemented in a hypothetical real-time system.

In a real-time scenario when the intelligent system predicts the upcoming action, we would like to answer the question when it is possible to predict the action before it happens. For this purposes, we analyzed sequences of one fixation and two fixations prior to the action, as demonstrated in Figure 4, and evaluated how predictive power changes further from the mouse click. The motivation here is to predict the upcoming action as soon as possible so that the adaptive system can proactively respond.

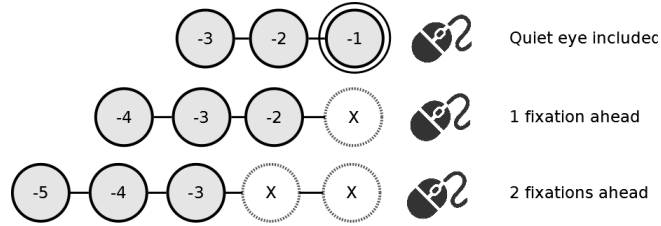


Fig. 4: Analysed datasets: the dataset with the quiet eye fixation (QE), 1-fixation and 2-fixations ahead of the action. The quiet eye fixation (double-circle) is the last fixation before the action. The empty circles indicate 1 and 2 excluded fixations.

2.4 Feature sets

Each sequence of gaze fixations was encoded into a feature vector represented by gaze events: fixation duration, distance between fixations, saccade duration, saccade orientation, saccade velocity and saccade acceleration. Each gaze event was described by statistical parameters (mean, median, variance, standard deviation, first, last) and ratio-based parameters (ratio of first vs. last fixation). All together each sequence was represented by the 54 gaze features. Datasets of 2-fixation based sequences were represented by 11 gaze features (fixation duration: mean, standard deviation, variance, sum, first, last, ratio of first vs. last, fixation distance, saccade duration, saccade direction and saccade velocity) as there were lacking enough gaze events for other statistical parameters.

We balanced datasets of action and non-action feature vectors. All available action-related sequences were included and non-action feature vectors were randomly sub-sampled. Since the amount of extracted actions slightly differed across datasets (within 10-50 samples), all datasets were evened out to 2500 action and 2500 non-action vectors. The balanced setting allowed to examine optimal setup for classifier training and for cross-study comparison. In real interaction, the proportion of user actions is more imbalanced, the original dataset contained over 2500 actions (12.5%) compared to 17500 non-actions. The search for the optimal class weights for the imbalanced datasets is out of scope of this work.

2.5 Classification framework

The classification in this study builds on the baseline prediction framework proposed by [5]. A nested parameter grid search employs Support Vector Machine (SVM) as a core classifier with an RBF kernel. Parameter search is the central process in model learning. In the two nested cross-validations (3 x 3-fold), the learning process estimated the most fitting hyperparameters for SVM-RBF (C, Gamma). The remaining settings, such as parameter grid search and SVM classifier, adhered to the classification standards employed in current machine learning studies. Using output SVM.C and SVM.Gamma, the model was built on training data (2/3 of the balanced dataset) and tested on the unseen data (1/3 of the balanced dataset).

3 Classification Results

To understand dependencies between context length (number of fixations in the sequence), context timing (omitting fixations prior to action) and gaze signal, we evaluated 6 sequence lengths (2 up to 7 fixations at once), and compared datasets with quiet eye (QE dataset) and two datasets prior to action; all together 18 datasets were classified using the classification framework introduced above. Here we report on classifier accuracy and Area Under the Curve (AUC), as the primary performance metrics, and training and testing AUCs for all datasets, as measures of classifier generalizability.

3.1 Context length and timing

When comparing length of the context, the datasets from longer fixation sequences performed better than the shorter ones. Figure 5 summarize classifier training performance for all timing variants and all sequence lengths. The best performance was reached with sequences of length 5 (accuracy=67.8%, AUC=0.735) and length 7 (accuracy=67.57%, AUC=0.745) in the QE dataset; the dataset 1-fixation-ahead performed best on sequence lengths 6 (accuracy=59.9%, AUC= 0.647) and the dataset 2-fixations-ahead of the action scored the best accuracy with length = 7 (accuracy = 60.6, AUC = 0.653) .

When comparing the effect of sequence length, the original datasets performed similar in all lengths with just slight deviations in accuracy and AUC; the difference between the best and worst performance was 2.6% in the accuracy measure and Δ AUC = 0.026. Both datasets prior to action revealed higher differences between the best and the worst performance than the QE dataset; the difference in the dataset 1-fixation-ahead of the action reached an accuracy of 6.43% (Δ AUC = 0.05) and sequences 2-fixations-ahead differed in the best and the worse accuracy of 9.2% (Δ AUC = 0.075).

The primary difference between QE and prior-to-action datasets was in missing quiet eye fixation [26], therefore we statistically compared durations of the quiet-eye fixation and two previous fixations.

Informally, we observed a moderate decrease in fixation duration towards the action. A one-way ANOVA showed a significant difference between the duration of three fixations: the QE-fixation (M = 243.96, SD=133.69), the previous fixation (M = 253.83, SD = 134.90) and the second last fixation (M = 294.10, SD = 158.92), $F(2, 3213) = 36.99$, $p < .001$. The mean duration of QE fixation was significantly shorter than second last fixation (Shapiro-Wilk test $p < 0.001$).

3.2 Generalizability of classifier

The classifier generalizability was evaluated using unseen data (1/3 of the whole dataset) in all datasets. Figure 6 captures the AUC rates for training and testing classifications using the QE and prior-to-action datasets. The classifier trained on the QE dataset proved stable performance in line with the training; the testing results outperformed the training performance (mean difference between testing

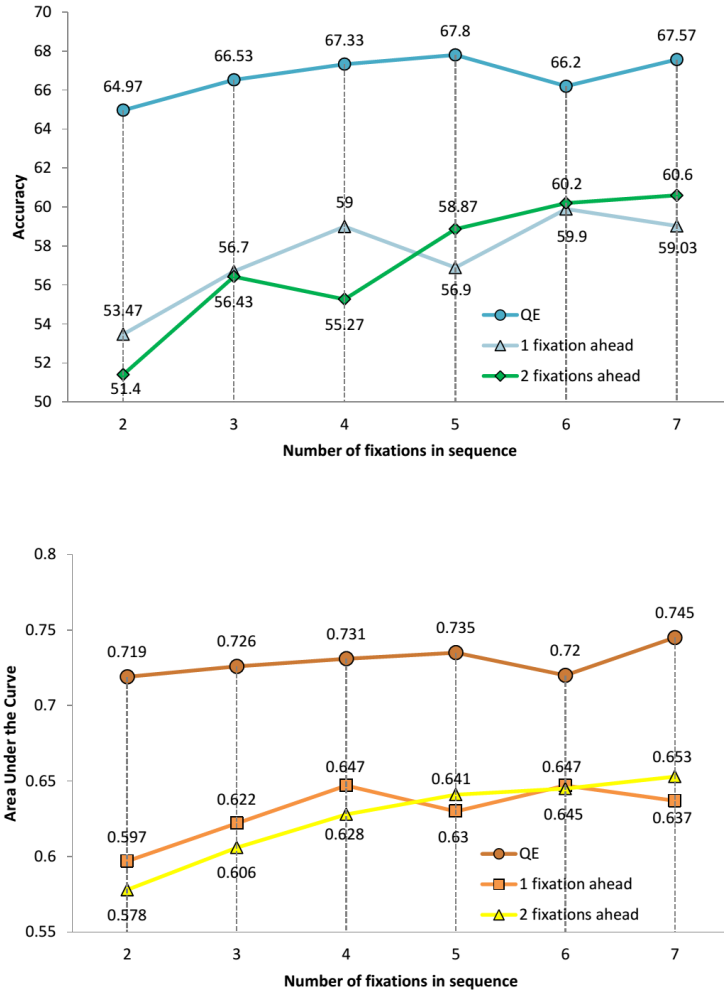


Fig. 5: Performance on datasets with different sequence lengths (number of fixations analyzed in the sequence). Accuracy (top) and Area Under the Curve (bottom).

and training AUC= 0.004, SD = 0.016), suggesting good classifier generalizability on the unseen datasets.

Differences between training and testing AUC for the prior-to-action datasets were also minimal (1-fixation prior to action: mean difference in AUC = 0.008, SD = 0.01) and (2-fixations prior to action: mean difference in AUC = -0.0018, SD = 0.01). The negative value of the difference indicates that training classifier AUC was on average better than the testing one.

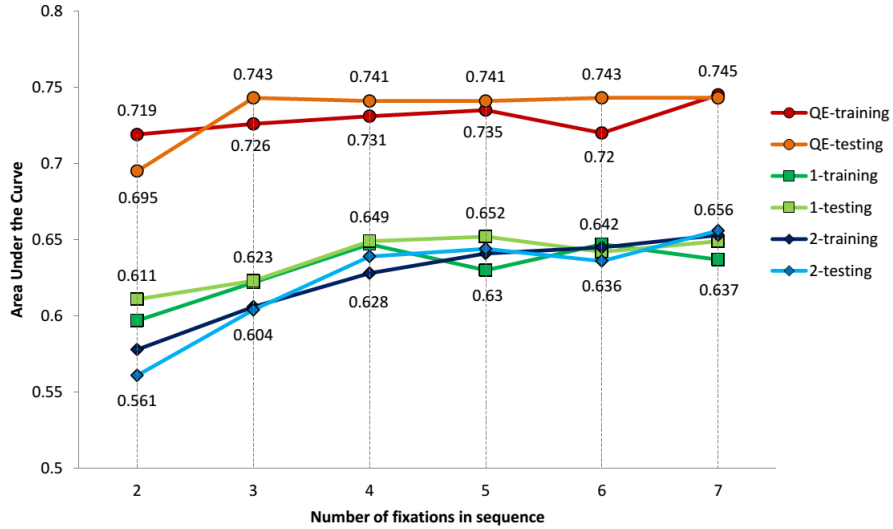


Fig. 6: Comparison of training and testing AUCs for all analyzed datasets.

4 Discussion

Detection of user actions remains a persistent challenge in intelligent system design. To tackle the action detection from gaze, we systematically modified context length (number of fixations) and timing (excluding quiet eye fixations and two fixations prior to the action), and classified the sequences with a standard SVM framework. We compared cross-validation and testing results to demonstrate generalizability of the approach. The main novelty is the analysis of the individual fixations and their contribution to the action detection performance.

4.1 Context length and timing

When comparing length of the context, the datasets from the longer fixation sequences performed better than the shorter ones. The best scores in terms of accuracy and AUC were received with 5 to 7 fixations in the sequence. However, the number of fixations did not affect the action detection when the quiet eye fixation was included. The systematic evaluation of the fixation sequences revealed minor improvements, the increasing performance was achieved generally with longer sequences. Omitting the quiet eye fixation from the analysis, on the other hand, decreased the classifier performance and highlighted the higher recognition rates in longer sequences.

Quiet eye fixation is the important part in movement programming of targeting tasks [29] and in our experiment, we observed how properties of quiet eye fixations contributed to action recognition. When we compared fixation dura-

tions of quiet eye fixations and previous two fixations, the quiet eye fixation was significantly shorter.

4.2 Applications of action detection and further considerations

In this work we investigated the links between interaction actions and user’s gaze, and how one can use such knowledge in the interaction design. The core of the work lies in computational evaluation of the connections between properties of the proactive gaze, the phases of actions and its manifestation. We envisage that such computational model, mediated by machine learning methods, will automatically detect the interaction actions from the stream of eye movements and inform the interactive system about user’s goals.

Our work can also extend the systems of classical activity recognition. For example, one explanation for the results obtained in this paper is that users were often multitasking. Thus, while performing the action -which is the ground-truth for our work- another planning activities may have taken the place. Therefore, the patterns of eye movements were not entirely constant and reflected at times other activities than action and planing. Here, the recent work on detection of interleaved activities [20] can be utilized.

5 Conclusion and future work

Reliable detection of user actions is a fundamental challenge in building intelligent interfaces that interact with the user in a human-like way. Based on the results of our study, it is safe to argue that eye movements reveal interactive actions and that eye-movement patterns around the interactive actions differ from other types of activities. We presented that processing eye movements as signals is a feasible way for detection of interaction actions. We adopted a pattern recognition system and applied it on the dataset from interactive problem solving. In order to evaluate its effects on the classification performance, we varied the contextual information and timing available for the decision, in terms of fixation counts and the analysis ahead of the action.

Our findings pointed out that the context length does not affect the action detection as much when the quiet eye fixation is part of the analyzed sequence. Systematic evaluation of the fixation sequences revealed minor improvements in the classification towards longer sequences. Omitting the quiet eye fixation from the analysis, on the other hand, decreased the classification performance and highlighted higher recognition rates with longer sequences.

Recent research in this domain employed classifiers trained on general population as descriptors and models of individual user behavior. Studies that would focus on personalized classification and performance differences when each participant presents own training and testing environment are however rare. In future work we will compare the performance of the existing system trained for the individual to the one trained for global behavioral data. It is essential that

future work will couple both personal and global models, and will assess the contribution of each.

Independently on the type of the adaptation style, the complex domains demand collections of reliable datasets with genuine samples of the ground-truth actions. While in this work the ground-truth was objective, eliciting higher-level actions, for instance an intention to influence a person, will require a careful methodological work.

References

1. Bailey, B.P., 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), 21:1–21:28 (2008)
2. Bartels, M., Marshall, S.P.: Measuring cognitive workload across different eye tracking hardware platforms. In: *Proc. of the Symposium on Eye Tracking Research and Applications*. pp. 161–164. ETRA '12, ACM (2012)
3. Bednarik, R., Eivazi, S., Vrzakova, H.: A computational approach for prediction of problem-solving behavior using support vector machines and eye-tracking data. In: Nakano, Y.I., Conati, C., Bader, T. (eds.) *Eye Gaze in Intelligent User Interfaces*, pp. 111–134. Springer (2013)
4. Bednarik, R., Gowases, T., Tukiainen, M.: Gaze interaction enhances problem solving: Effects of dwell-time based, gaze-augmented, and mouse interaction on problem-solving strategies and user experience. *J. of Eye Movement Research* 3(1), 1–10 (2009)
5. Bednarik, R., Vrzakova, H., Hradis, M.: What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In: *Proc. of the Symposium on Eye Tracking Research and Applications*. pp. 83–90. ETRA '12, ACM (2012)
6. Bixler, R., DMello, S.: Toward fully automated person-independent detection of mind wandering. In: *User Modeling, Adaptation, and Personalization*, pp. 37–48. Springer (2014)
7. Bondareva, D., Conati, C., Feyzi-Behnagh, R., Harley, J.M., Azevedo, R., Bouchet, F.: Inferring learning from gaze data during interaction with an environment to support self-regulated learning. In: *Artificial Intelligence in Education*. pp. 229–238. Springer (2013)
8. Bulling, A., Gellersen, H.: Toward mobile eye-based human-computer interaction. *Pervasive Computing, IEEE* 9(4), 8–12 (2010)
9. Bulling, A., Roggen, D., Troster, G.: What's in the eyes for context-awareness? *Pervasive Computing, IEEE* 10(2), 48–57 (2011)
10. Eivazi, S., Bednarik, R.: Inferring problem solving strategies using eye-tracking: system description and evaluation. In: *Proc. of the 10th Koli Calling Int. Conference on Computing Education Research*. pp. 55–61. ACM (2010)
11. Eivazi, S., Bednarik, R., Tukiainen, M., von und zu Fraunberg, M., Leinonen, V., Jääskeläinen, J.E.: Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. In: *Proc. of the Symposium on Eye Tracking Research and Applications*. pp. 377–380. ETRA '12, ACM (2012)
12. Flanagan, J.R., Johansson, R.S.: Action plans used in action observation. *Nature* 424(6950), 769–771 (2003)
13. Heckhausen, H., Beckmann, J.: Intentional action and action slips. *Psychological Review* 97(1), 36–48 (1990)

14. Hradis, M., Eivazi, S., Bednarik, R.: Voice activity detection from gaze in video mediated communication. In: Proc. of the Symposium on Eye Tracking Research and Applications. pp. 329–332. ETRA '12, ACM (2012)
15. Jacob, R.J.K., 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, pp. 573–605. Elsevier Science (2003)
16. Just, M.A., Carpenter, P.A.: A theory of reading: From eye fixations to comprehension. *Psychological review* 87, 329–354 (1980)
17. Kardan, S., Conati, C.: Comparing and combining eye gaze and interface actions for determining user learning with an interactive simulation. In: User Modeling, Adaptation, and Personalization, pp. 215–227. Springer (2013)
18. Kautz, H.A., Allen, J.F.: Generalized plan recognition. In: AAAI. vol. 86, pp. 32–37 (1986)
19. Memmert, D.: Pay attention! a review of visual attentional expertise in sport. *Int. Review of Sport and Exercise Psychology* 2(2), 119–138 (2009)
20. Modayil, J., Bai, T., Kautz, H.: Improving the recognition of interleaved activities. In: Proc. of the 10th int. conf. on Ubiquitous computing. pp. 40–43. UbiComp '08, ACM (2008)
21. Norman, D.A.: *The Design of Everyday Things*. Basic Books, New York (2002)
22. Prabhu, P., V., Prabhu, G., V.: *Handbook of Human-Computer Interaction Chapter 22 Human Error and User-Interface Design*. Elsevier Science B. V. (1997)
23. Rayner, K.: Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124(3), 372 (1998)
24. Simola, J., Salojärvi, J., Kojo, I.: Using hidden markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research* 9(4), 237 – 251 (2008)
25. Steichen, B., Carenini, G., Conati, C.: User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In: Proc. of the 2013 int. conf. on Intelligent user interfaces. pp. 317–328. ACM (2013)
26. Vickers, J.N.: Visual control when aiming at a far target. *J. of Experimental Psychology: Human Perception and Performance* 22(2), 342 (1996)
27. Vrzakova, H., Bednarik, R.: Eyecloud: Cloud computing for pervasive eye-tracking. In: PETMEI 2013, 3rd Int. Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction (2013)
28. Vrzakova, H., Bednarik, R.: Fast and comprehensive extension to intention prediction from gaze. In: IUI 2013 Workshop on Interacting with Smart Objects (2013)
29. Williams, A.M., Singer, R.N., Frehlich, S.G.: Quiet eye duration, expertise, and task complexity in near and far aiming tasks. *Journal of Motor Behavior* 34(2), 197–207 (2002)