What do you want to do next:
A novel approach for intent prediction in gaze-based interaction

Roman Bednarik*
School of Computing
University of Eastern Finland

Hana Vrzakova†
School of Computing
University of Eastern Finland

Michal Hradis‡
Faculty of Information Technology
Brno University of Technology

Abstract

Interaction intent prediction and the Midas touch have been a long-standing challenge for eye-tracking researchers and users of gaze-based interaction. Inspired by machine learning approaches in biometric person authentication, we developed and tested an offline framework for task-independent prediction of interaction intents. We describe the principles of the method, the features extracted, normalization methods, and evaluation metrics. We systematically evaluated the proposed approach on an example dataset of gaze-augmented problem-solving sessions. We present results of three normalization methods, different feature sets and fusion of multiple feature types. Our results show that accuracy of up to 76% can be achieved with Area Under Curve around 80%. We discuss the possibility of applying the results for an online system capable of interaction intent prediction.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Input devices and strategies;

Keywords: gaze-based interaction, Midas touch, machine learning, activity detection

1 Introduction

> At first, it is empowering to be able simply to look at what you want and have it happen, rather than having to look at it (as you would anyway) and then point and click it with the mouse or otherwise issue a command. Before long, though, it becomes like the Midas Touch. Everywhere you look, another command is activated; you cannot look anywhere without issuing a command. The challenge in building a useful eye tracker interface is to avoid the Midas Touch problem. [Jacob 1991]

Reducing and at best a complete eradication of the Midas Touch problem is one of the central themes in applied eye-tracking research [Jacob and Karn 2003; Istance et al. 2008]. Every gaze-based interface, for instance an eye-typing software, has to separate the intentional activation of a command from other behaviors, in order to avoid the Midas touch problem.

In this paper, we propose a system capable of predicting whether a user of a gaze-based interface intends to issue a command. Our hypothesis is that one way to solve the challenge is to approach the problem from a machine learning perspective. Such an approach allows processing of multidimensional high-frequency eye-tracking data, while allowing for efficient and effective implementation. Before such a system is implemented, however, we need to find the most suitable set of features that describe human intents and methods for their processing.

2 Related research

Fortune-tellers master the study of pupil dilation and eye movements for so called mind-reading. Using well chosen questions, they infer the deep intentions we are desperate to hide. Humans cannot easily resist mind-reading, since the control of pupil dilations and eye movements is as possible as the control of the heartbeat. If intent prediction, based on this phenomena of mind-reading, is within the human abilities, can an artificial intelligence with the same reasoning skills be created?

Pupillary responses have long been known for reflecting the sensory, behavioral, cognitive and emotional states of a user. In particular, these pupillary reflex dilations, also called psychosensory reflexes [Kahneman 1973], are defined as the pupillary reaction to outer sensory signals, such as touch, visual or audio stimuli. Many internal mental processes and states also trigger pupil dilations, and they typically consist of emotions, attention and other mental workload processes.

In the present work we explore the types of pupillary responses related to internal states. The size of such pupil dilations is distinctly smaller than the size of pupillary changes caused by natural reflexes, see [Beatty and Lucero-Wagoner 2000], on the other hand, their well-established correlations with cognitively intensive processes motivate their adoption in applied studies of human-computer interaction. Although pupillary responses are a mixture of reactions to lighting, closeness of an object and cognitive load, they are considered as markers of internal states. This makes them an attractive means for understanding of human thoughts, emotions and actions in eye-tracking research.

An example of such a pupillary response dependent marker is *Movement-related pupillary response* (MRPR), which was previously investigated in relation to self-triggered finger flexes by Richer and Beatty [1985]. The physical effort for pressing a button and the movement complexity resulted in significant differences in the pupil diameter. The differences appeared approximately 1.5 seconds before and peaked 0.5 seconds after the performed movement [Richer and Beatty 1985]. The typical pattern of pupil dilation related to MRPR is shown in Figure 1.

Previous research also discovered links between mental workload during problem solving and pupil diameter. According to Beatty [1982], pupillary responses are task-dependent and correspond to levels of task difficulty, and could possibly be used as a marker of intelligence. Bailey et al. [2008] and Iqbal et al. [2004] presented a framework for detecting task boundaries based on pupil dilations as a measure of cognitive load.

Furthermore, Klingner [2010] found a link between mental effort

*e-mail:roman.bednarik@uef.fi
†e-mail:hanav@uef.fi
‡e-mail:ihradis@fit.vutbr.cz

ETRA 2012, Santa Barbara, CA, March 28 – 30, 2012.
© 2012 ACM 978-1-4503-1225-7/12/0003 $10.00

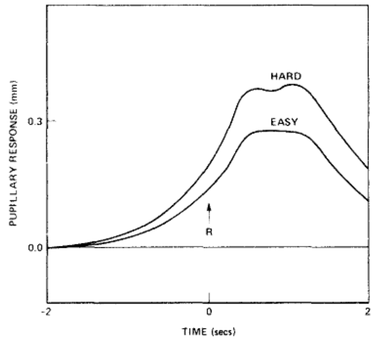

Figure 1: *The typical course of movement-related pupillary response. At time 0 a movement of a finger was initiated. Figure from [Richer and Beatty 1985].*

and pupillary responses during map reading and searching tasks. Looking up for a given locality caused significant differences in pupil size compared to legend reading, whereas symbol memorizing caused decrease in the pupil diameter.

Besides pupillary responses, other eye-tracking measures were previously shown to correlate with mental workload and internal processing. A longer mean fixation duration, for example, can be attributed to attention rising and depth of required processing, as shown in the case of web browsing [Jacob and Karn 2003], problem-solving [Eivazi and Bednarik 2011], or higher information priority during skim reading [Duggan and Payne 2011].

Any application of the findings reviewed above faces the central problem of insufficient understanding of the interplay between low level signals of eye movements, the processed eye movement measures and, finally, the high-level behavioral and interactive events such as the intention to issue a command. Here, we propose to investigate this link using machine learning methods.

Machine learning (ML) and classification approaches have previously been adapted in eye-tracking research to automatically process the great volumes of eye-movement data. The core aim of all attempts is to find some computational structure that describes a link between the low-level raw data and high-level behavioral units.

Simola et al. [2008] achieved about 60% prediction accuracy when inferring in which of three states a user can be during information search tasks. Other instances of a machine learning approach for eye-tracking data analysis have been reported by Bednarik et al. [2005] and Kinnunen et al. [2010]. The authors have applied a state-of-the-art biometric person authentication system based on either traditional signal processing methods such as PCA or FFT. Using eye movement velocity and pupil size, or a histogram of the velocity and gaze direction, the authors achieved identification rates of 60% or equal error rates of 29%, respectively. Further attempts at eye-tracking based biometrics, using machine learning methods, have been reported too, e.g. by Kasprowski and Ober [2004].

Recently, Eivazi and Bednarik [2011] applied a Support Vector Machine (SVM) based approach for learning and classifying cognitive states during problem-solving. The authors employed seven combined eye-tracking features and achieved an accuracy of approximately 53% on a five-class classification problem.

The eye movement challenge held in 2005, reported by Salojärvi [2005] can be considered as the only systematic dwelling into the issues of method and feature selection for ML analysis of eye movement data. The purpose of the systematic evaluation proposed by Salojärvi [2005] was to achieve implicit feedback for a recommender system and to improve the query generation. The authors later described an implementation of the proposal by Ajanki et al. [2009] and expanded the idea into inferring object relevance in dynamic scenes, as described by Kandemir [2010].

Based on previous research, we assume human interaction intentions can be inferred from eye movements. Although it seems that the previous research offers machine learning as a potential solution for this challenge, little is known about the suitable methods, approaches and features that would allow for efficient and practically applicable intention prediction. Thus, the research questions we answer using an experimentation approach are:

- What eye movement features describe best the differences in gaze behavior during intentional and non-intentional interaction?
- What are the effects of gaze data normalization techniques on the prediction outcomes?
- Is the approach computationally efficient and reliable enough for real-time applications and practical purposes?

3 A machine learning system for interaction intent prediction

The machine learning framework applied here can be considered standard. The main challenge in application of ML to eye-tracking data lies in finding an appropriate sets of features and effective methods [Eivazi and Bednarik 2011]. We propose feature sets to be built from fixation-related measures, saccadic movements and pupil diameter signals. A variable length time window is then represented by a constant length feature vector accompanied by a label. The label encodes whether the particular time frame corresponds to an observed interaction intention or not. After extracting feature sets, machine learning is employed for building models able to distinguish between the intentional and non-intentional feature vectors, by means of training and parameter optimization. In the following sections, we describe the building blocks in a more detail.

3.1 Data labelling and ground truth

We define the concept of '*intention*' as an internal sequence initiated by the user, resulting in an interface action. An interface action can be, for instance, pressing a button, moving a window, or selecting a piece of text. In an eye-typing interface, for example, a user would have an intention to type the character. Using one of the currently employed approaches, she would then typically spend a prolonged time on the intended letter and the gaze-based interface would generate a command to print the letter [Majaranta and Räihä 2002]. At other times, however, the user would not have the intention to type and we define these instances as *non-intention*.

3.2 Features and extraction

The extraction of intentional and non-intentional sequences is based on an adaptive window shifted through consequential fixations. We selected this approach –against the shift by a constant number of data samples in the time domain– to decrease the possibility of having two interaction intentions falling into one window[1]. Thus, we extract the sequences by selecting a specific number of fixations and respective gaze data samples contained within the range of the selection. Not only is the fixation-bound approach more convenient,

[1] We assume that the rate of interaction intentions is smaller than the rate of fixations.

it also is computationally effective for a future real-time processing, since it works with less input data concurrently.

Figure 2 presents a hypothetical example of intentional and non-intentional fixation sequences with one fixation overlap.

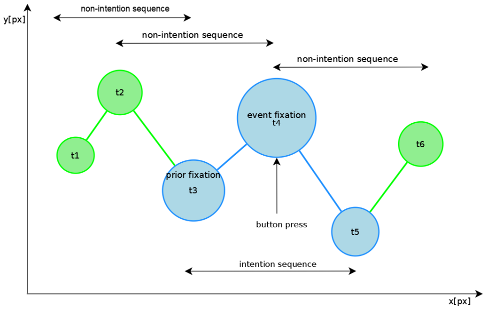

Figure 2: *Intention and non-intention three-fixation sequences. The intentional sequence (blue) frames the event fixation with another two fixations (prior- and post-event fixation), whereas the non-intentional sequences (green) are extracted with one fixation overlap.*

Due to the lack of standardization in eye-tracking metrics, the choice of the best describing features often leans to exploratory investigations [Jacob and Karn 2003]. We employed two approaches to feature computation, in which all gathered samples, bounded by the observed fixations, serve as an input for feature extraction. The first type of features are gained from positions and durations of fixations and saccadic movements, as suggested by Jacob and Karn [2003]. These features are summarized in Tables 1 and 2. We investigated 13 features, based on fixation duration, saccade length, and saccadic velocity and acceleration.

The second approach focuses on pupillary dilations within the processed fixations. To compensate for variations in fixation lengths, we processed raw pupil diameter signals by employing signal normalization, pupil alignment and linear warping so that all sequences were unified to the same length, as introduced by Klingner [2010]. Our choice of features, based on pupillary responses, is unique to the field of eye-tracking data analysis, and therefore in the remainder of the section we motivate their use for inferring internal cognitive state prediction.

Table 1: *Eye movements features computed from fixations*

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

Features based on uniform length pupillary responses, illustrated in Table 3, are based on signal processing methods using power spectrum and cepstrum. The power spectrum is able to reveal slow and fast dynamic changes in pupil diameter [Kinnunen et al. 2010], while the cepstrum shows a rate of change among the different spectrum groups [Bogert et al. 1963]. As a marker of pupil signal stability, the first and the second differences and their one-dimensional histograms are calculated.

Table 2: *Eye movements features computed from saccades*

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

Another feature within our set employs the Dynamic Time Warping algorithm (DTW) [Sakoe and Chiba 1990]. The algorithm calculates curve similarity between randomly chosen reference sets and the tested pupillary responses. The resulted distance serves as an input feature for machine learning. The reason of concentrating on this feature lies in a hypothesis that intentional fixation sequences differ from the non-intentional ones in the shape of the pupillary response curve, similarly as shown by Richer and Beatty [1985] for MRPR.

The last extracted feature uses the average percentage change in pupil size (APCPS) [Iqbal et al. 2004], computed as an average of Formula 1, since we assume intention generation may cause an increase in pupil diameter. Such increases may not be observable in non-intentional sequences.

$$PCPS = \frac{\frac{X-\mu}{\sigma}}{\mu} \tag{1}$$

Table 3: *Features derived from pupillary responses*

Pupillary response features	Feature description
Spectrum	Power spectrum of the pupil diameter signal
Cepstrum	Power cepstrum of the pupil diameter signal
First difference	Histogram of the first differences
Second difference	Histogram of the second differences
DTW distance	Degree of pupil signal similarity to reference set
APCPS	Average percentage change in pupil size (APCPS) over the fixation sequence

3.3 Classifier, training and optimization

The features defined above serve as an input for the prediction model, aiming to classify the feature vector as an intentional or non-intentional movement. For the purpose of the binary classification, Support Vector Machine (SVM) is considered the best off-the-shelf classifier. SVM finds optimal (maximum margin) separating hyperplanes in the original feature space or in an expanded feature space when a kernel is used [Cortes and Vapnik 1995]. SVM provides a satisfactory trade-off between classification performance and computational demands. The ease-of-use, task-independence and existing ready-to-use method implementations, lead us to select the SVM as the classifier core.

A structure of the SVM based framework used in this study is illustrated in Figure 3. It consists of input and normalization operations, realized by Z-transform [Oppenheim et al. 1999], and a nested cross-validation block responsible for iterative subset splitting [Mierswa et al. 2006] that serves as an input for model training and testing exclusively.

We employ a six-fold outer (estimation of classification performance) and six-fold inner (estimation of optimal SVM hyperparameters) cross-validation. In each cross-validation iteration, the a data set is divided into training and testing part using stratified sampling which guarantees that the ratio of classes is preserved in the split subsets. SVM hyperparameters are selected using a grid search algorithm.

The grid search exhaustively seeks for optimal parameters, defined by intervals and steps, then selects one combination of parameters, a grid point, and evaluates the learning performance. In this way, the parameter combinations with the highest performance is found and examined on the entire training set. These parameters are then applied as settings of the resulting prediction model [Chang and Lin 2011].

Grid search parameters are evaluated by applying them on a testing subset and by measuring their performance. Typically, classification accuracy is used as a evaluation metric. However, the challenges of relying only on classification accuracy were already mentioned by Provost et al. [1998] and thus, employing other more convenient metrics, such as the Area Under the Curve (AUC) [Egan 1975] is generally recommended.

A problem of unbalanced dataset with disproportional number of vectors in one class (rare class) is overcome by the afore mentioned AUC evaluation. In addition, setting of the class weights in SVMs to increase the importance of the rare class solves the challenge of the ratio between positive and negative vectors. A similar effect is achieved by setting a threshold specifying costs for class missclassification. Thereby, assignment of the rare class representative as the incorrect class has a higher penalty than the opposite case and thus, leads to a better balanced classification [Mierswa et al. 2006].

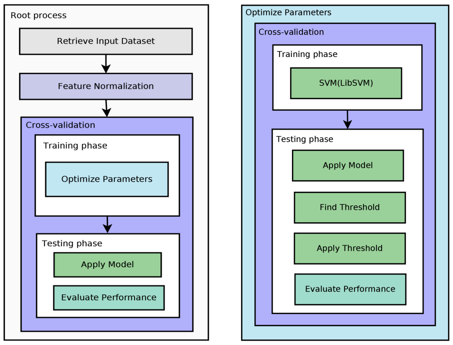

Figure 3: *Root and parameter optimization processes of the prediction model. The model main blocks are held in the root process, while the parameter optimization steps are given within the cross-validation.*

3.4 Normalization

Pupillary responses, when used as a feature, face limitations due to specific illumination conditions and participants' biological as well as emotional characteristics [Beatty and Lucero-Wagoner 2000]. Hence, an absolute size of the pupil diameter, as measured by the eye-tracker, requires normalization that transforms the pupil sizes to the same range for different participants or sessions.

It is also a well-known fact that normalization affects the classification and prediction results. In the instance of eye-tracking, it is not yet known how the features should be normalized, and we thus evaluated three different methods, namely:

- Baseline subtraction - A baseline is defined as a mean of pupil diameters within the observed sequence or dataset. Normalized pupillary responses are achieved by subtracting the baseline from the input pupillary responses [Klingner 2010].
- Z-score - A normalized signal is gained using baseline subtraction and dividing a partial result by the signal standard deviation [Hupé et al. 2009].
- Percent changes in pupil size (PCPS) - A normalization is performed by baseline subtraction from the input pupil signal, subsequently a partial result is divided by the baseline [Beatty 1982].

Above mentioned normalizations were employed in two variants, differing in the unit of normalization (*participant*-wise normalization vs. *sequence*-wise normalization). The effects of the methods are illustrated in Figure 4.

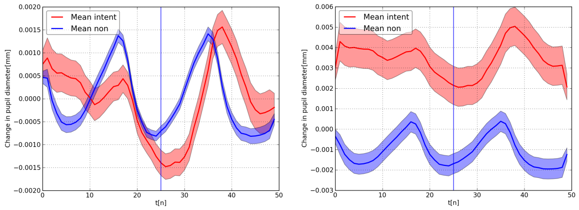

Figure 4: *Pupillary responses normalized over sequence (left) and dataset (right) by PCPS normalization. The mean representation of intentional (red) and non-intentional (blue) sequences allows a visual comparison how the chosen normalization unit influences the pupillary response curves.*

In addition to pupillary response, we also normalized fixation and saccadic features to have a zero mean and unit variance. This type of scaling assures that all features are equally important to the classifier and that values of the optimal hyperparameters are in a common range.

4 Method and Case dataset

The method proposed here is based on extracting gaze-related features that distinguish intentions from non-intentions during interaction with a gaze-based interface. In contrast to previous research that employed only fixations and saccades, this work also aims to utilize pupil dilations during gaze-based interaction.

The dataset was originally collected for another purpose, as is described by Bednarik et al. [2009]; the original study examined the effects of interaction modality on problem-solving activities. Here we briefly explore the characteristics of the user interface used and the study settings.

4.1 User interface and interaction technique

The problem-solving interface is shown in Figure 5. The goal of the puzzle was to arrange the tiles into the target configuration, shown at the bottom left corner. A tile could only be moved if there was an empty adjacent space. A gaze-augmented interaction was designed so that when the gaze is fixated on a valid tile and a button is pressed, the tile is moved to the empty space.

Compared to, for example, purely dwell-based interaction where navigation and action would be a part of one modality, the separation between the navigation for selecting a tile and issuing a command allows us to more accurately distinguish between action (intention) and non-action. We also did not include the data from a simple mouse-based interaction, because the rate of mouse-clicking was rather high [Bednarik et al. 2009], and thus little data was available for training the models for the non-interaction part of the data.

The button press in the gaze-augmented condition also sets the boundaries for the proposed fixation sequence extraction. The corresponding sequence of eye tracking data is related to this event. In this application of the proposed framework, we study gaze behavior during two fixations before the event and one fixation after it, to prevent event overlapping.

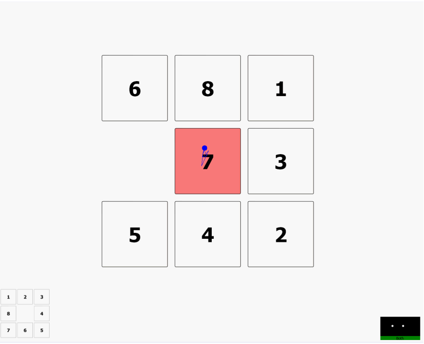

Figure 5: *The user interface of the 8-tile slide puzzle. User is gazing to the center and could see the target solution (bottom left) and the status of the eye-tracking (bottom right). Gaze-path is not visible to the user.*

4.2 Design, participants, and procedure

The original study had a mixed subject design, with one between factor (an interaction modality, three levels: mouse, gaze-augmented, dwell-time) and one within factor (trial, three levels). In this report we employ only one third of the original experimental data, i.e., we only included the data from the gaze-augmented condition.

The reference dataset consisted of 13 participants with normal or corrected-to-normal vision, in the age range from 21 to 53. A recording of one participant had to be eliminated from the dataset since the calibration was poorly performed and caused significant delay between user interaction and interface responses.

The recordings were conducted in a quiet usability lab with constant illumination. Participants first conducted a warm-up session to get familiar with the interaction condition, and then performed a block of three trials. The duration of a trial was not limited, and on average participants solved the problem in 218 seconds [Bednarik et al. 2009].

4.3 Apparatus

The dataset was recorded using a table-mounted Tobii ET 1750 eye tracker. The eye tracker was set at a sample rate 50Hz binocularly and, reported by the manufacturer, can achieve an accuracy of around 0.5 degree of visual angle. Tobii's ClearView [Tobii Technology AB 2008] with default settings for fixation identification was employed as the eye movement recording tool, and to capture the mouse button presses.

4.4 Data analysis

Th whole fixation set was obtained from raw data by filtering using the following settings: fixation radius 30 pixels and minimal fixation duration was set to 100ms; data were averaged over both eyes. Thereafter, fixation sequences extraction and features set preparation was implemented using a set of custom Python scripts, using the contribution of Numpy, Scipy and Matplotlib libraries [Jones et al. 2001–]. Preliminary analyses were performed using the R statistical tool. The design of the prediction model and the procedure of parameter optimisation was run in RapidMiner 5.0, an open source solution supporting machine learning and data mining [Mierswa et al. 2006]. The prediction model (SVM) was realized as a RapidMiner operator with support of LibSVM [Chang and Lin 2011]. The experiments were performed at a server computer, containing eight processors, 2.93GHz Intel Xeon X3470, and 10GB RAM, running Ubuntu Linux distribution.

The approach presented in the previous section was applied for the experimental analysis of the case dataset, and we varied the normalization technique and the type of features. A preparation phase of feature extraction and normalization were performed using Python scripts. During script processing, each dataset contained all three trials and the intentional sequences were extracted around the occurrences of button-press events. The rest of data were labeled as non-intentions. The window size was set to three fixations and the ratio of overlapping as two fixations.

The proposed architecture of the prediction model was employed in three nested loops. The main loop contained a nested 6-fold cross-validation and employed stratified sampling. The upper layer performed model training and testing, while the lower level implemented model parameter optimization. The chosen parameters consisted of SVM kernel parameters *C*, in range of $\langle 10, 900 \rangle$ in 4 steps with the logarithmic scale, and γ, in range of $\langle 1E^{-8}, 5E^{-4} \rangle$ also in 4 steps with the logarithmic scale. The output optimization parameter was the Area Under the Curve to be maximized, so that the classifier considered the unbalanced distribution of labels.

5 Results

Altogether, the systematic evaluation comprised 41 experiments, with a total running times over 180 hours.

5.1 Descriptive results

Table 4 presents the counts of sequences containing an intention to interact compared to all other sequences. The complete dataset is not balanced, as the non-intentional sequences were nearly 7 times more frequent. A baseline majority classifier would thus have an accuracy of 87% but precision of 0%.

Figure 6, 7 and 8 show three examples of the feature distributions. It is interesting to observe the shape of the distributions. First, the histograms of the measures related to fixations and saccades resemble normal distributions symmetric around a peak, however only for the sequences belonging to the interaction intentions. All other

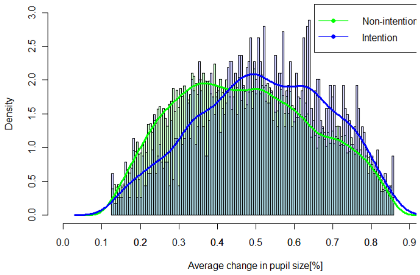

Figure 6: *Histogram of average percentage of pupillary responses (APCPS)*

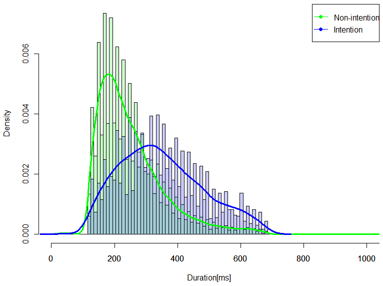

Figure 7: *Histogram of mean fixation durations.*

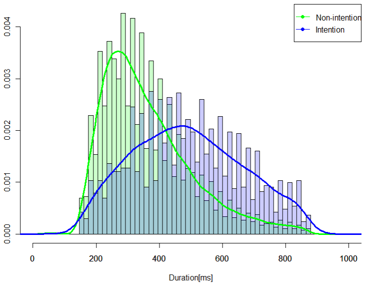

Figure 8: *Histogram of sums of saccade duration.*

Table 4: *Input dataset statistics*

Participant	Intention sequences	Non-intention sequences	Sum count
1	383	1797	2180
2	18	1082	1100
3	130	2030	2160
4	73	953	1026
5	305	1601	1906
6	379	1575	1954
7	352	1687	2039
8	318	974	1292
9	119	1145	1264
10	104	849	953
11	148	1666	1814
12	161	1390	1551
Total count	2490	16749	19239
Total count[%]	12.94	87.06	100

data points seem to generate a relatively skewed distribution. The differences in the distributions present a potential for the machine learning methods. On the other hand, the pupil-size related vectors seem to show a better overlap, thus they potentially are more difficult to distinguish.

5.2 Classification results

The experiments were based on a binary classification and thus, performance of the classifier was expressed as a confusion matrix. In the present case, the accuracy of the prediction model was not a reliable measure, since the dataset was unbalanced and consequently, the accuracy would not reflect this problem (the overall accuracy could be high even though one class was misclassified [Egan 1975]). Thus, in addition to accuracy (ACC) we employed the receiver operating characteristic graph (ROC) and Area Under Curve (AUC) for the classifier evaluation.

Table 5 presents an overview of the main results, listing the best achieved classification performances for each of the main combinations of the feature sets and normalization methods. The best performing combination (fixations, saccades, DTW, and APCPS features) achieved an AUC of 0.807, with false negative rate of 30.5% and false positive rate of 24.0%. An exhaustive listing of all results is presented in an online appendix[2].

Table 5: *Overall classification results. Note: F + S refers to a fusion of all fixation and saccade features. Type of normalization unit is indicated in parentheses, seq. = fixation sequence, set = whole participant dataset. C and γ are optimized SVM parameters.*

Features	Normalization	ACC	AUC	C	γ
DTW	Z-score (seq.)	0.351	0.562	30.8	1.00E-08
1st DIFF	Z-score (seq.)	0.432	0.572	900.0	5.00E-04
2nd DIFF	PCPS (seq.)	0.510	0.540	292.2	5.00E-04
APCPS	PCPS (seq.)	0.517	0.577	94.9	1.36E-05
F + S	Z-transform (set)	0.751	0.799	900.0	5.00E-04
F + S + DIFF2	Z-transform (set), PCPS(set)	0.743	0.799	900.0	5.00E-04
F + S + DTW	Z-transform (set), Z-score (seq.)	0.753	0.800	900.0	5.00E-04
F + S + APCPS	Z-transform (set), PCPS (seq.)	0.758	0.806	900.0	5.00E-04
F + S + DTW + APCPS	Z-transform(set), Z-score (seq.), PCPS (seq.)	0.759	0.807	900.0	5.00E-04

Table 6 shows the comparison of the normalization methods on the classification performance based on pupillary response features. A 3 x 2 (normalization method x unit of normalization) ANOVA did not discover any significant effects, implying that there were no differences between the methods.

Table 6: *Classification outcomes with respect to normalization methods and pupil-related features.*

	Normalization					
	Baseline subt.		Z-score		PCPS	
Unit of normalization	Set	Seq.	Set	Seq.	Set	Seq.
AUC	0.556	0.553	0.552	0.566	0.569	0.565
Accuracy	0.304	0.366	0.368	0.430	0.368	0.430
FPR	0.776	0.692	0.673	0.701	0.676	0.594
FNR	0.161	0.245	0.230	0.238	0.304	0.382
TPR	0.839	0.755	0.770	0.762	0.696	0.618
TNR	0.224	0.308	0.327	0.299	0.324	0.406

6 Discussion

The original hypothesis, which drives the presented research, expected that intents were observable in human gaze as specific patterns of eye movements, and therefore, these patterns could possibly be parametrized by a set of features in order to build a prediction

[2] http://cs.uef.fi/~hanav/appendix_etra_2012.pdf

model. Special expectations were put on the use of pupillary dilations as a promising parameter. In previous research the changes in pupil size were linked to task-dependent actions and higher cognitive load. Another motivation was to conduct a ML comparison of pupillary responses and fixation- and saccade-based features on a single dataset, an attempt not previously reported.

The findings, however, confirmed this theory only partially. The features, computed from fixation and saccade positions and durations, were shown to be well-discriminative in the cases of intentional and non-intentional eye movement sequences. The AUC of the prediction model reached up to 0.8 which is a fairly good classification quality. On the other hand, the models based on the pupillary responses did not achieve equally good performance; the most solid AUC was measured around 0.6 and obtained by the histogram of the second difference. In addition, the pupillary response spectrum and cepstrum could not have been evaluated at all since the classifier was unable to be trained.

The fusion of fixation-, saccade- and pupillary dilation features tended to slightly improve the performance results. A numerical difference between the single-feature results and fusion-based classification was around 1%, which can be interpreted as an error of measurements. In sum, the pupillary dilations, as applied in our method, did not show large discriminative characteristics, as was extracted in the study of task-evoked pupillary responses; we thus conclude that our results do not match [Klingner 2010]'s findings that investigated cognitive load and pupillary responses only. A likely explanation is that the previous research rests on carefully collected, laboratory controlled, hand-picked and averaged pupillary data evaluated by a visual analysis, while on contrary our datasets contain voluminous instances of all naturally occurring pupillary responses.

Concerning the effects of normalization type, our results do not provide any grounds for recommending a generally well performing normalization technique. Our results, however, show that certain normalizations work better for certain features. This finding partly disapproves the previous research that recommends using one type of normalization [Hupé et al. 2009].

Another explanation for the low performance of pupillary responses is that the task-evoked responses are minor (in magnitude) compared to the responses induced by light and focus reflex. However, our experiments were carried out in a laboratory under stable lighting conditions and the interactive application did not evoked light contrasting situations, such as shade and contrast changes, which would lead to abnormal pupil size changes. Admissibly, the physiological changes could have happened if the participants looked outside of the screen and thus, pupils would have to adjust to the background surrounding. On the other hand, the number of such voluntary gazes away from the user interface was markedly lower than the overall number of fixations on the screen. Thereby, the error caused by off-screen gazes was likely not influencing the overall results.

In gaze-base interaction research, where the ability of the interface to recognize a command is central, the user is often required to indicate his intention by prolonged looking at an interaction element. These previous approaches rest on the eye-mind hypothesis, in the sense that direct intentional looking is associated with an interface action. We took a different, multidimensional perspective that can in future contribute to a more seamless interaction with user interfaces. Our contribution is the first one suggesting a leap away from the eye-mind approach for recognition of an incoming command from the user.

Another contribution of this study is a detailed description and adoption of a machine learning framework for the purposes of eye-tracking data analysis. This exposure allows others to validate our proposal with relative simplicity. We believe that improvements to the presented approach, and thus advances to the ML methods for eye-tracking data analysis, can be achieved by conducting a benchmarking comparison. We plan to release the dataset to public and invite others to present their classification approaches and results.

6.1 Online implementation

The results above provide a solid evidence for our hypothesis that there is an automatically detectable difference between intentional interaction and non-intentional looking. Such a system, if used alone, would still perform inadequately in real-time, by raising false alarms.

It is however easy to imagine that the presented prediction of intention could be used in conjunction with other methods for gaze-based interaction. For example, the interfaces based on dwell-time could benefit from the knowledge whether the user intends to issue a command. In particular, in the event of initiating the dwell-time timer, a gaze-based interface could adaptively set the duration of the dwell-time, depending on the information of intention prediction.

Conclusions

This study took a previously unexplored approach into an analysis of user intentions from a large eye movement dataset. We focused on the prediction of events related to interaction intention, by evaluating two important aspects of the ML methodology: feature set and normalization effects. We undertook this exploration using an unbalanced dataset, a situation naturally occurring in interacting with user interfaces.

The combination of the feature set presented here, the data processing methods and the classification approach do not currently fulfill the conditions for real-time classification. With the measured AUC, a hypothetical gaze-based interface would miss a low number of intended commands, and perhaps more importantly, could inadequately propose false events.

This study opens a new front in the battle with intent recognition based on eye movement data. The future steps include more studies on the features; for example, the DTW distance is a promising direction. Creating a pre-computed bank of action-dependent pupillary responses could improve the classification performance, as well as an adaptation for a background model and for personal differences. Such sequences, corresponding to distinct events, attitudes and cognitive load, can lead to higher classification performance since the reference dataset would be chosen on an a priori knowledge rather than by random selection, as it was done in this experiments. We also plan to investigate empirically, what the bounds of tolerable error related to intent prediction are, in order to set the requirements for online classification of intention.

References

AJANKI, A., HARDOON, D. R., KASKI, S., PUOLAMÄKI, K., AND SHAWE-TAYLOR, J. 2009. Can eyes reveal interest? implicit queries from gaze patterns. *User Model. User-Adapt. Interact. 19*, 4, 307–339.

BAILEY, B. P., AND IQBAL, S. T. 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction 14*, 4 (Jan.), 1–28.

Beatty, J., and Lucero-Wagoner, B. 2000. *The pupillary system.* Cambridge University Press, ch. 6.

Beatty, J. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol Bull 91*, 2 (Mar.), 276–292.

Bednarik, R., Kinnunen, T., Mihaila, A., and Fränti, P. 2005. Eye-movements as a biometric. In *14th Scandinavian Conference on Image Analysis, SCIA 2005*, Springer, 780–789.

Bednarik, R., Gowases, T., and Tukiainen, M. 2009. 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, 3–10.

Bogert, B., Healy, M., and Tukey, J. 1963. The quefrency alanysis of time series for echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. In *Proc. Symp. on Time Series Analysis*, 209–243.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning 20*, 273–297. 10.1007/BF00994018.

Duggan, G. B., and Payne, S. J. 2011. Skim reading by satisficing: evidence from eye tracking. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, ACM, New York, NY, USA, CHI '11, 1141–1150.

Egan, J. P. 1975. *Signal Detection Theory and ROC Analysis.* Academic Press.

Eivazi, S., and Bednarik, R. 2011. Predicting Problem-Solving Behavior and Performance Levels from Visual Attention Data. In *2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction at ACM IUI 2011*, 9–16.

Hupé, J.-M., Lamirel, C., and Lorenceau, J. 2009. Pupil dynamics during bistable motion perception. *Journal of vision 9*, 7 (Jan.), 10.

Iqbal, S. T., Zheng, X. S., and Bailey, B. P. 2004. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI '04 extended abstracts on Human factors in computing systems*, ACM, New York, NY, USA, CHI EA '04, 1477–1480.

Istance, H., Bates, R., Hyrskykari, A., and Vickers, S. 2008. Snap clutch, a moded approach to solving the midas touch problem. In *Proceedings of the 2008 symposium on Eye tracking research &#38; applications*, ACM, New York, NY, USA, ETRA '08, 221–228.

Jacob, R. J. K., and Karn, K. S. 2003. 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, 573–605.

Jacob, R. J. K. 1991. The Use of Eye Movements in Interaction Techniques: What You Look At is What You Get. *Human-Computer Interaction 9*, 152–169.

Jones, E., Oliphant, T., Peterson, P., et al., 2001–. SciPy: Open source scientific tools for Python.

Kahneman, D. 1973. *Attention and effort*. Englewood Cliffs, Nj: Prentice-Hall.

Kandemir, M., Saarinen, V.-M., and Kaski, S. 2010. Inferring object relevance from gaze in dynamic scenes. In *ETRA '10 Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 105–108.

Kasprowski, P., and Ober, J. 2004. Eye movements in biometrics. In *Biometric Authentication*, D. Maltoni and A. Jain, Eds., vol. 3087 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 248–258.

Kinnunen, T., Sedlak, F., and Bednarik, R. 2010. Towards task-independent person authentication using eye movement signals. In *ETRA '10 Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, 187–190.

Klingner, J. 2010. Fixation-aligned pupillary response averaging. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10*, ACM Press, New York, New York, USA, vol. 1, 275.

Majaranta, P., and Räihä, K.-J. 2002. Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, ACM, New York, NY, USA, ETRA '02, 15–22.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. 2006. Yale: Rapid prototyping for complex data mining tasks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, 935–940.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. 1999. *Discrete-time signal processing (2nd ed.).* Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Provost, F. J., Fawcett, T., and Kohavi, R. 1998. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '98, 445–453.

Richer, F., and Beatty, J. 1985. Pupillary dilations in movement preparation and execution. *Psychophysiology 22*, 2, 204–207.

Sakoe, H., and Chiba, S. 1990. Readings in speech recognition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ch. Dynamic programming algorithm optimization for spoken word recognition, 159–165.

Salojärvi, J., Puolamäki, K., Simola, J., Kovanen, L., Kojo, I., and Kaski, S. 2005. Inferring relevance from eye movements: Feature extraction. In *Helsinki University of Technology*, No, 2005.

Simola, J., Salojärvi, J., and Kojo, I. 2008. Using hidden markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research 9*, 4, 237 – 251.

Tobii Technology AB, 2008. Clearview 2.7 eye gaze analysis software. `http://www.tobii.com/scientific_research/`.