Accurate Measurements of Pointing Performance from In Situ Observations

Accurate Measurements of Pointing Performance from In Situ Observations

Krzysztof Z. Gajos, Katharina Reinecke and Charles Herrmann
Harvard School of Engineering and Applied Sciences
Cambridge, MA 02138, USA
{kgajos,reinecke,iherrmann}@seas.harvard.edu

ABSTRACT

We present a method for obtaining lab-quality measurements of pointing performance from unobtrusive observations of natural in situ interactions. Specifically, we have developed a set of user-independent classifiers for discriminating between deliberate, targeted mouse pointer movements and those movements that were affected by any extraneous factors. To develop and validate these classifiers, we developed logging software to unobtrusively record pointer trajectories as participants naturally interacted with their computers over the course of several weeks. Each participant also performed a set of pointing tasks in a formal study set-up. For each movement, we computed a set of measures capturing nuances of the trajectory and the speed, acceleration, and jerk profiles. Treating the observations from the formal study as positive examples of deliberate, targeted movements and the in situ observations as unlabeled data with an unknown mix of deliberate and distracted interactions, we used a recent advance in machine learning to develop the classifiers. Our results show that, on four distinct metrics, the data collected in-situ and filtered with our classifiers closely matches the results obtained from the formal experiment.

Author Keywords

Pointing, motor performance, machine learning, in situ studies

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation: User Interfaces—*Evaluation/methodology*;

General Terms

Human Factors, Experimentation

INTRODUCTION

We present a method for obtaining lab-quality measurements of pointing performance from unobtrusive observations of natural in situ interactions. In controlled laboratory studies, participants are instructed to focus entirely on the task at hand and proceed as quickly as possible while keeping error rates minimal. The measurements collected in such studies reflect primarily the motor abilities of the participants and the properties of the input device used. In contrast, in naturalistic settings people's pointing performance is also affected by a number of cognitive, perceptual, or environmental distractions such as deciding what task to perform next, looking for the right user interface element, or watching TV. Thus, data obtained from in situ observations of natural pointing interactions differs substantially from data collected in laboratory settings [2]. Yet, if lab-quality data could be collected in situ, it might enable measuring longitudinal fluctuations in a user's motor performance (e.g., due to medication, fatigue, or progression of a disease) [11], automatic detection of pointing problems and the need for assistive technology [10], automatic adaptation of user interfaces to the changes in the user's motor abilities [7], and more realistic assessments of novel input devices.

To enable such lab-quality measurements of pointing performance from in situ observations we have developed a set of user-independent classifiers for discriminating between deliberate, targeted mouse pointer movements and those movements that were affected by any extraneous factors. As illustrated in Figure 1, these classifiers can be used to effectively filter the naturalistic data such that the filtered in situ data is nearly indistinguishable from the data obtained from the same person through a formal experimental set-up.

To develop and validate our approach we developed logging software to unobtrusively record mouse pointer trajectories from 18 participants as they naturally interacted with their computers over the course of several weeks. Each participant also performed a set of pointing tasks using a formal study set-up. For each movement, we computed a set of measures capturing properties of the movement trajectory as well as the speed, acceleration, and jerk profiles. Treating the observations from the formal study as positive examples of deliberate, targeted movements and the in situ observations as unlabeled data with an unknown mix of deliberate and distracted interactions, we used a recent advance in machine learning [4] to develop user-independent classifiers capable of discriminating between trajectories of deliberate and distracted mouse pointer movements.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

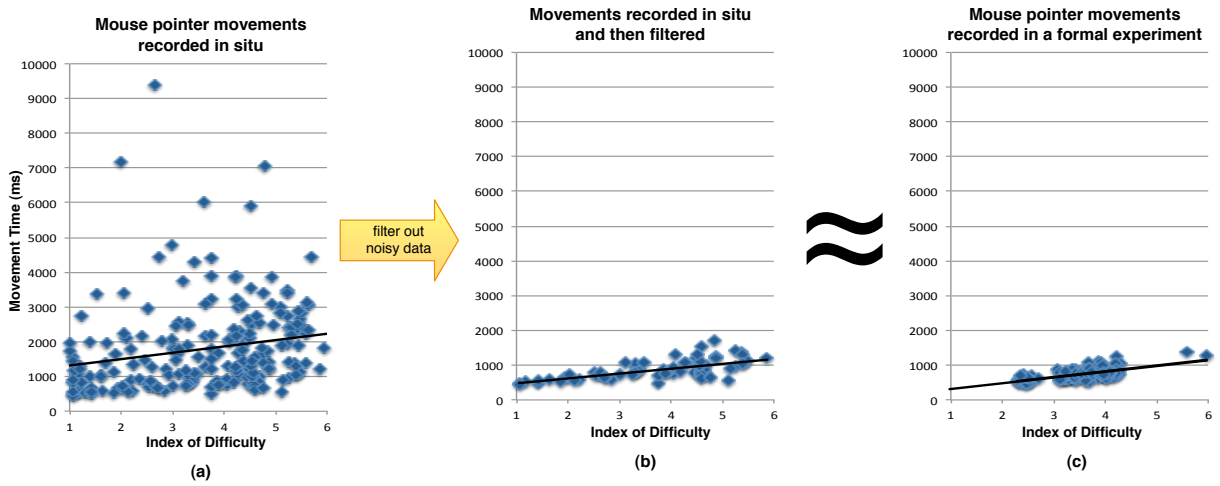

Figure 1. (a) Mouse pointer movements collected in situ from one participant by unobtrusively observing his interactions with the computer. (b) A subset of these movements classified by our system as deliberate targeted interactions; the distribution of these movements, as represented by the Fitts's law parameters computed from them, is nearly indistinguishable from (c) the movements collected from the same participant in a formal experiment. Trend lines are shown in black.

We evaluated our approach on four different tasks — estimating the mean of a measurement, conducting tests for statistically significant differences, estimating Fitts' throughput, and building Fitts' law models to generate movement time predictions — and the results obtained from data collected in situ and filtered with our classifier closely matched the results obtained from a formal experiment.

The main contribution of this paper is the development and the evaluation of a machine learning–based method for discriminating between mouse pointer trajectories that came from deliberate, targeted movements and those that were produced when the person was distracted or not focused on the task. We also make available on our web site our data set of over 7800 mouse pointer trajectories, our source code, and the user-independent classifiers trained on these data.

PRIOR RESEARCH

Unobtrusive observations of user behavior have previously been used to characterize naturally occurring mouse pointer interactions [2]. The analysis of these data confirmed that naturally occurring pointing interactions have different properties from those observed in controlled laboratory studies. Meanwhile, Hurst and colleagues demonstrated that laboratory studies are insufficient for properly characterizing the motor performance of people with motor impairments because the motor abilities of these users tend to vary substantially over time due to medication, fatigue, and changes in the underlying medical condition [11] — longitudinal in situ measurements are needed instead. Hurst and colleagues also demonstrated that lab-quality pointing performance data could be used to automatically detect the certain kinds of motor impairments and to predict the benefit of particular assistive technologies [10] and, again, argued that their methods would be most useful if the data could be collected unobtrusively.

The SUPPLE system [6] uses models of an individual's motor performance to automatically generate user interfaces optimized for that user's unique motor abilities. Currently, lab-quality measurements are needed to build appropriate models of users' motor abilities. The data collection process used by SUPPLE can take up to an hour for people with moderate to severe impairments to complete. Such significant upfront effort will deter many users from adopting ability-based adaptations, especially if the effort has to be repeated each time a person's abilities fluctuate or their computing environment changes.

Unobtrusively collected behavioral measures — capturing temporal patterns in scrolling, mouse movements, clicks, text entry, and periods of apparent inactivity — have also been used to predict the reliability of contributions by workers on the Amazon Mechanical Turk crowdsourcing platform [19]. Our work could provide another mechanism for assessing participant effort (and therefore data quality) for tasks that involve pointing as part of the overall interaction.

The very recent work on the Input Observer [5] is aimed at developing a logging and analysis system for both keyboard- and mouse pointer–based interactions. It uses a conservative heuristic for identifying parts of movements that can be used to compute Fitts' model parameters characterizing a particular individual's motor performance. If combined with Input Observer, our work may enable more efficient use of the collected data.

GOALS AND APPROACH

Our high level goal, illustrated in Figure 1, is to obtain results from in situ observations that are of the same quality as the results obtained from controlled laboratory experiments.

Our approach is to relax requirements for how the data are collected and instead filter the observations by identifying deliberate, targeted mouse pointer movements and separating

them from those that were affected by any extraneous factors such as cognitive effort related to a difficult choice, visual search for the correct target, lack of effort, or an environmental distraction.

Our approach hinges on the assumption — which our results confirm — that deliberate, targeted mouse pointer movements can be distinguished from "noisy" ones by a careful analysis of their trajectories as well as the speed, acceleration and jerk profiles.

To collect examples of deliberate, targeted mouse pointer trajectories, we asked each participant to perform a set of pointing tasks based on the ISO 9241-9 standard [12]. To obtain a broad and representative sample of mouse pointer trajectories affected by extraneous factors, we asked participants to install a Firefox browser plugin on their computers for a period of several weeks to record the trajectories of the mouse movements performed while participants were interacting naturally with their computers.

We evaluate the approach by examining its reliability on four representative tasks:

- estimating a mean of a directly measured variable for each participant (we use movement time divided by Fitts' Index of Difficulty);
- testing for statistically significant differences in the values of those means across participants;
- estimating the per-participant value of the Index of Performance;
- computing parameters of the Fitts' law model for each participant and using the model to make movement time predictions for a range of tasks.

We briefly review Fitts' law, and the related concepts of Index of Difficulty and Index of Performance in the Data Processing section.

Definitions

For ease of exposition, we will refer to the data collected from the formal pointing experiment test as *experimental* data or movements. *Natural* data or movements will denote the mouse pointer trajectories collected unobtrusively with the browser extension. Finally, our approach will classify the natural movements into *deliberate* and *distracted*.

DATA COLLECTION

Participants

18 participants (10 male, 8 female; aged 19 to 64) successfully completed all parts of the study. Participants were recruited nationally from the broad population through advertisements delivered via email or posted on various community portals. All participants were regular computer users. Almost all participants reported using a touchpad, a mouse, or a combination of the two as their primary input devices. Only one participant used a tablet. These data are summarized in Table 1.

Participant	Pointing device used typically	Age	Gender
P02	touchpad	21	m
P03	mouse	52	f
P17	touchpad	30	m
P18	touchpad	32	f
P22	touchpad	20	f
P24	mouse/touchpad	24	m
P25	touchpad	30	m
P30	touchpad	19	f
P31	touchpad	22	m
P33	touchpad, sometimes mouse	27	f
P40	mouse	42	m
P41	mouse	29	m
P43	touchpad	20	f
P44	mouse	64	f
P48	mouse/touchpad	52	f
P49	mouse	52	m
P60	tablet	31	m
P63	mouse	32	m

Table 1. The Participants.

No participants reported any major impairments or other factors that might affect their computer usage except for P48 who reported slightly reduced visual acuity (partially corrected with an anti-glare screen) and P44 whose mouse acted "sticky."

Participants received monetary compensation for taking part in the study.

Apparatus

We have built two data collection tools: a web-based formal experimental tool based on the ISO 9241-9 standard [12] (to collect the experimental data) and a Firefox browser extension to record trajectories of the mouse pointer movements that occurred naturally within the browser content window. The formal experimental tasks were presented as sets of targets arranged in a circular pattern. The next target to be clicked was highlighted in yellow. A number of task sets were presented to each user varying in target sizes (10, 15, 25, 40, and 60 pixels) and the distances between successive targets (three distances dependent on the size of each participant's screen). Up to 10 clicks were collected per condition and the conditions were presented in a random order.

The data recorded with both tools included all mouse movement events and clicks. The tools also recorded when the mouse left and re-entered the browser content window so that we could identify those movements for which the trajectory data was incomplete (our software had no access to mouse events triggered outside of the browser content window). For the explicit diagnostic tasks, we recorded whether a click was a hit or a miss. For data captured with the Firefox extension, we were able to identify only a subset of meaningful interactors (those that were marked with tags such as A, INPUT, SELECT, etc., and those with appropriate ARIAA [22] tags).

When a click occurred on such a target, we recorded the target type, its dimensions and location. When a click occurred on any other part of the page, we also recorded it, but because we could not tell if the element clicked was a valid interactor, we could not reliably determine if the click was a miss or a deliberate interaction.

To maintain participants' privacy, our software did not collect the URLs or the content of the pages visited by the participants, nor did it capture any information entered by the users on the pages. All data collected on participants' computers were identified by a study code and were periodically automatically transmitted to our server over an encrypted and authenticated connection.

Procedure

Interested participants responded to our advertisements by visiting the study web site where they were presented with a brief overview of the study purpose and procedures, followed by an informed consent form. Participants who consented were asked to provide basic contact information and were given instructions on how to download and install the Firefox extension. They were then contacted by a researcher, who conducted a phone interview to collect basic demographic information and who confirmed the study procedures and helped resolve any technical issues.

The unobtrusive recording of a participant's mouse pointer movements began the moment a participant consented to participate in the study and installed the Firefox extension.

Participants were instructed to do the formal experiment at as quick a pace as possible and at a time when they would not be distracted. They were told that they could take a break from those tasks at any point they needed to. Our software automatically resumed where the participant had left off. Participants were encouraged to complete the diagnostic tasks as soon as possible, but were free to do so any time before the completion of the study.

After 4 weeks, participants were contacted again by a researcher with instructions on how to uninstall the Firefox extension.

DATA PROCESSING

Extracting Individual Mouse Movements

We parsed the collected sequences of mouse pointer events to extract individual *movements.* We defined a movement as a sequence of mouse move events starting immediately after a click that ended a previous movement or after a pause of at least 1 second. Valid movements ended with a click on a recognized interactive element. Because our software could not identify all interactors on every page, if a click occurred outside one of the recognized interactors, we could not reliably determine if such a click was a mistake or an intentional action — we did not include such movements in the analysis. We also discarded movements during which the mouse pointer left the browser content window, the user switched to a different tab, or a page load event was registered.

Because we could not always obtain the exact position of a clicked target, we calculated movement distance as the distance from the location of the first movement event to the location of the final click.

We used the following formulation of the Fitts' law to compute any Fitts' law–related measures:

$$\text{Movement Time} = a + b * \log_2\left(\frac{A}{W} + 1\right)$$

Where A is the movement *amplitude,* or distance, and W stands for the target size. Constants a and b are estimated from the data. The logarithm term is referred to as the *Index of Difficulty,* or ID and is measured in bits. The reciprocal of b is called the Index of Performance and is measured in bits per second, and is occasionally used as an informative diagnostic measure for comparing performance of devices or people [15, 21]. We used the smaller of the target width and height as the target size [17] (more sophisticated approaches are available [1, 8]; we chose the min model because it provides reasonable accuracy with minimal complexity).

Data Cleaning

We removed extreme outliers from the data collected during the formal experimental tasks. From each participant's data we removed those movements where the Movement Time divided by ID was two standard deviations or more away from that participant's mean. This resulted in removal of between 3 and 11 movements per participant (between 2% and 7% of each participant's data).

Filtering Trajectories

To enable robust computation of trajectory-related statistics, we first smoothed the pointer trajectories to remove any sampling-related artifacts. We started by resampling pointer position trajectories at regular time intervals (10 ms resulting in a 100 Hz sampling rate) and we subsequently filtered them with a 30 Hz low-pass NER filter [14]. To compute successive derivatives (speed, acceleration, and jerk), we used a differentiating filter [14]. Once the derivatives were computed, we passed position, speed, acceleration, and jerk time series again through a 7 Hz low-pass filter. This overall process has been used previously by a number of researchers (e.g., [21, 13, 18, 20]) to ensure that low-level sampling-related artifacts are removed, while significant phenomena are preserved.

Computing Movement Features

In addition to movement time, we recorded a number of other performance-related measures (or features) for each movement. These measures have been described, illustrated, and motivated more extensively by the researchers who originally proposed them, but we briefly summarize them below for completeness. In defining some of the measures, we will refer to the *task axis,* which is the straight line passing through the end points of the movement. The following measures refer directly to the movement trajectory:

Length to Distance Ratio [2] is the ratio of the length of movement's overall trajectory to the straight line distance between the movement's end points.

Number of Task Axis Crossings [16] captures how many times the pointer crossed the task axis.

Movement Direction Changes [16]—if the movement trajectory were rotated such that both end points were on the x-axis, this measure captures the number of times the y-component of the movement direction changed sign (i.e., it captures direction changes toward/away from the task axis).

Orthogonal Direction Changes [16]—as above, but for the x-component of the movement direction (i.e., it captures direction changes toward/away from the target).

Movement error [16] is computed as the average absolute distance of the mouse pointer from the task axis.

Movement variability [16] captures the standard deviation of the distance of the points on the trajectory from the task axis. Intuitively, it captures the level of "wiggliness" of the movement.

Speed, acceleration, and jerk profiles. Following [21], we also computed several measures of the speed, acceleration, and jerk profiles:

- the number of peaks
- the number of x-axis crossings
- time from the start of the movement to the first peak
- the relative time (with respect to the total movement duration) from the start of the movement to the first peak
- time from the start of the movement to the last peak
- the relative time to the last peak
- time from the start of the movement to the maximum peak
- the relative time to the maximum peak

Finally, we computed correlations between task properties (distance, target size, and the Index of Difficulty) and the values of each of the features. We found that 4 features (movement time, and the time of the last peak for each of speed, acceleration, and jerk) correlated substantially with the Index of Difficulty. Because our goal is to use these features to identify differences in participant performance that are independent of the task, we replaced each of these 4 features with a transformed version by dividing the original by the Index of Difficulty. Lastly, we added Movement Time divided by the movement distance, and Movement Time divided by a logarithm of the movement distance to the list of features.

Thus, for each movement we computed a total of 34 features.

Overview of the Data

In the end, we extracted a total of 7847 usable mouse pointer trajectories (2762 from the formal experimental tasks and 5085 from the browser extension). The browser extension was much easier to develop than recording tools that integrate directly with the operating systems, but it resulted in a relatively low data collection rate: A large fraction of the recorded movements were affected by the mouse pointer leaving the content window, slow page loads, or tab switches. Some sites also prevented recapturing of mouse click events. The extended data collection period ensured, however, that sufficient data were collected for the purposes of this study.

As a diagnostic check, we computed the Index of Performance for each participant's experimental movements. The results ranged from 4.1 bits/second for our oldest participant to 10.5 bits/second for the youngest, with the mean of 6.9 bits/second (a number that is consistent with the results from other studies, e.g., [21]).

CLASSIFYING MOUSE MOVEMENTS

Given the collected data, we aimed to develop a classifier capable of discriminating between mouse pointer trajectories corresponding to deliberate and distracted interactions. Note, however, that our data did not match the assumptions of typical classification approaches: while the experimental data could be assumed to contain mostly examples of deliberate movements, the natural data was an unlabeled mix of deliberate and distracted movements, and we could not even justify making assumptions about the relative proportions of the two types of movements. While typical classification approaches assume that all data are labeled and that all classes are represented in roughly equal proportions, our data set contained some positive labeled examples (the experimental data) and the remaining data had no labels at all (the natural data). Many other classification problems that arise in HCI would be most naturally treated as learning from positive and unlabeled examples (e.g., a user quickly inspects a large number of images and marks a small number as interesting — the remaining images will be more correctly treated as unlabeled rather than uninteresting). We therefore briefly summarize the state of the art in machine learning to support other researchers in adopting these techniques in their work.

The problem of learning from positive and unlabeled examples has been addressed by several machine learning researchers. Some early approaches involved using heuristics to identify an initial set of likely negative examples among the unlabeled data and then falling back on better understood semi-supervised learning techniques (e.g., [23]). Other approaches make assumptions about the relative proportions of the positive and negative examples in the unlabeled data (e.g. [3]). All of these approaches can be very sensitive to the accuracy of the initial guesses. A more recent approach [4] overcomes the limitations of the earlier work. This approach builds on the observation that if the labeled examples are drawn completely at random from the pool of all positive instances in the training data, then, informally:

$$p(\text{instance x is positive}) = \frac{p(\text{instance x would be labeled})}{\text{fraction of positive instances that are labeled}}$$

In other words, to learn from positive and unlabeled examples, we can first fall back on a more traditional classification problem where we assume that all unlabeled examples are in fact negative. Then $p(\text{instance x would be labeled})$ is a probability estimate that a mainstream classifier would return if trained on a data set where all unlabeled examples are temporarily given a negative label. This probability then needs to be adjusted by a constant specific to the data set: the fraction of positive instances that are labeled. This quantity cannot be

observed directly, but the reader will find three methods for estimating it in Elkan and Noto's paper [4].

This overall method is well-motivated theoretically and is also easy to implement because it relies on existing classification methods for estimating the probabilities. The main requirement is that the underlying mainstream classification method produces well-calibrated probabilities. Many of the standard classification techniques used in the HCI community, such as decision trees, support vector machines, or naive Bayes, do not satisfy this requirement. Instead, logistic regression is an appropriate choice, and this is what we use in the work presented here.

EVALUATION PROCEDURE

Evaluation Measures

Intuitively, our goal is to measure how "alike" the data obtained from the formal experiment and the natural data filtered with our approach are. Formal measures of distribution similarity are unlikely to be illuminating, however. We therefore turn to four functional metrics that reflect common purposes for which such data might be used. Using the experimental data as the base line for all error computations, we compute per-participant error rates for the following primary measures:

Mean Movement Time divided by the Index of Difficulty (MT/ID). For each participant, we compute the difference in the estimates of the mean value of Movement Time divided by the Index of Difficulty (MT/ID) between the experimental data and the natural data classified as deliberate.

Tests for statistically significant differences. For each pair of participants, we test whether the difference in their mean values of MT/ID is significant. We use a t-test at $\alpha = .01$ and we log-transform the data before applying the test to correct for the non-normal distribution. We perform these tests separately for both data sets (the experimental data set and the data set being evaluated). To make the comparison between the two data sets meaningful, we randomly subsample the larger data set such that the two data sets have the same number of measurements for each participant. To make the results less susceptible to chance, we repeat each test 100 times redoing the resampling each time. We consider the outcome of the test as significant if more than half of the repetitions returned $p \leq .01$. We limited the testing to only those pairs where at least 50 data points were available for each participant.

Movement time predictions of Fitts' models. For each participant, we estimate the parameters of the Fitts' models separately from the experimental data and the natural data classified as deliberate. We compute the relative difference between the predictions of the two models with the model derived from the experimental data serving as the baseline of the comparison. We perform this comparison for a set of Index of Difficulty values ranging from 2 to 7 at intervals of 0.2. Because the models may cross (i.e., each model may produce higher estimates for a part of the range of the IDs considered), we calculate the mean *absolute* relative difference in the predictions.

Mean Index of Performance. For each user, we compute the relative difference between the estimates of the Index of Performance derived from the experimental data and the natural data classified as deliberate. This is the most sensitive of our measures because it is estimated with the least squares method, which is particularly sensitive to outliers for extreme values of ID.

In addition to the above error measures, we also compute three more informative statistics:

Standard deviation of the MT/ID. While the errors in the estimates of the mean value of MT/ID capture the systematic errors in the measurements, standard deviation captures the variance in the measurements and provides another indication of the data quality. We compute the mean per-user standard deviation.

Recall. Recall captures the fraction of the experimental movements that the classifier identifies as deliberate. Assuming that our participants completed the formal experimental tasks conscientiously and without distraction, it is desirable for this number to be close to 100%.

Fraction of natural movements classified as deliberate. This measure provides some insight into how generalizable the model used by the classifier is—a very small fraction of natural movements classified as deliberate may indicate that the classifier was overfitting to the data from the training set. This number is also informative for estimating how many more in situ observations would need to be collected than in a controlled experiment for similar quality of the results.

Cross-validation Procedure

We used a per-participant cross-validation procedure for evaluation: for any configuration tested, we trained the classifier on data obtained from all but one participant. We applied the resulting classifier to the data from the remaining participant to identify the deliberate movements and we computed the values of all of the evaluation measures on those filtered data. This procedure ensured that the results reflect the capability of this approach to make accurate classification decisions on data collected from previously unseen users.

Feature Selection Procedure

The choice of features impacts the ability of machine learning algorithms to generalize robustly. We have identified over 30 potential features (discussed in the Data Processing section), but which of them are actually useful for discriminating between deliberate and distracted movements?

We used a variant of the stepwise discriminant analysis as our feature selection algorithm. Starting with an initial (potentially empty) set of features to consider, our algorithm proceeded in the following steps: First, it removed one feature at a time from the current set of features. If removing any one feature resulted in an improvement, the best of such reduced feature sets was kept as the current best and the process was restarted with this set as the starting point. If removing features did not improve performance, the algorithm attempted

	Unfiltered Natural Movements	*Approach 1* Filtered	*Approach 2* Normalized & Filtered	*Approach 3* Filtered, Target-Agnostic Features Only	*Approach 4* Normalized & Filtered, Target-Agnostic Features Only
Properties and assumptions					
Data requirements	n/a	movement trajectory + target size	movement trajectory + target size + a set of trajectories representative of the person's natural behavior	movement trajectory	movement trajectory + a set of trajectories representative of the person's natural behavior
Movement Time divided by Index of Difficulty					
Mean absolute error	172.10%	22.00%	19.05%	35.14%	25.12%
min	*70.78%*	*-33.20%*	*-20.41%*	*-3.24%*	*-4.43%*
max	*236.92%*	*51.42%*	*45.32%*	*77.97%*	*63.93%*
Mean per-participant standard deviation (value for experimental data: 62)	488.54	52.90	53.71	103.58	89.16
Statistical tests					
Fraction incorrect	49.67%	41.18%	16.67%		
chance of not detecting correctly a statistically significant difference if there is one	76.62%	51.33%	42.86%	6.82%	24.14%
chance that a statistically significant difference will be reported as significant but in opposite direction	*5.26%*	*0.00%*	*0.00%*	*0.00%*	*1.72%*
chance of detecting a statistically significant difference	22.37%	12.50%	10.34%	55.74%	59.57%
Movement time predictions of the Fitts' models					
Mean absolute error	149.38%	21.85%	19.42%	35.90%	21.48%
min	*67.18%*	*9.39%*	*0.79%*	*3.62%*	*2.67%*
max	*219.10%*	*54.96%*	*44.07%*	*75.21%*	*49.56%*
Index of Performance					
Mean absolute error	-27.27%	20.00%	16.82%	24.07%	29.08%
min	*-68.17%*	*-48.32%*	*-47.55%*	*-55.02%*	*-25.57%*
max	*78.05%*	*63.61%*	*38.01%*	*22.55%*	*81.04%*
Recall					
Mean recall on experimental data	n/a	89.03%	88.75%	84.98%	88.70%
min	*n/a*	*24.78%*	*64.60%*	*26.55%*	*69.12%*
max	*n/a*	*99.35%*	*99.45%*	*98.70%*	*98.35%*
Mean fraction of natural movements classified as deliberate	n/a	24.65%	23.12%	30.65%	24.77%
min	*n/a*	*3.90%*	*7.25%*	*11.11%*	*12.95%*
max	*n/a*	*58.70%*	*36.12%*	*55.80%*	*35.68%*

Table 2. Summary of the results. Errors are reported as percentages of the values derived from the experimental data.

to extend the model by adding one feature (trying each of the unused features in turn). If any of such expanded feature sets resulted in an improvement, the process was restarted with the best of such expanded sets as the starting point. The process terminated when neither shrinking nor expanding the feature set resulted in an improvement. To reduce the chance that the final result would be a local minimum, the process was repeated with a number of different randomly chosen starting sets of features.

The numerical criterion used to evaluate the performance of classifiers trained on different subsets of the feature was the sum of the mean and worst-case values of the MT/ID error, error in the movement time prediction of the Fitts' models, and the error in the Index of Performance estimate.

RESULTS

In this section, we report on the performance of four variants of our approach, each making different assumptions about the type and quantity of the available data.

We begin by evaluating how informative the raw unfiltered natural data are for evaluating the quantities listed in the Evaluation Measures section. These and other results are summarized in Table 2. As expected, the natural data resulted in substantial overestimates of the mean value of MT/ID and much larger variance in the data. Next, nearly half of the pairwise statistical tests resulted in a conclusion different from what was obtained from experimental data: Not surprisingly, most of the errors were false negatives: nearly 77% of the statistically significant differences present in the experimental data were not detected in the natural data. Of larger concern are the false positives and the small fraction of tests that indicated the presence of a significant difference, but in the direction opposite to that observed in the experimental data. A larger number of natural observations may help correct the false negative errors, but the other two types of errors are likely due to the different distributions of "noise" in each participant's natural data and may not be correctable even if larger amounts of data were collected.

The natural data also resulted in Fitts' models that substantially differed in their predictions from the models calculated from the experimental data.

Index of Performance is the most indirect measure we collected from the data. The mean absolute difference between the estimates obtained from natural and experimental data is 27%. Closer inspection of the results shows a strong bias toward underestimates: for all but three participants the Index of Performance estimated from the natural data is lower than the ground truth.

Approach 1: Identifying Deliberate Movements

The first approach addresses the situation where movements have to be classified on-line, one-by-one. We assume that both the movement trajectory and the size of the target are available.

Approach 1 **Filtered**	*Approach 2* **Normalized & Filtered**	*Approach 3* **Filtered, Target-Agnostic Features Only**	*Approach 4* **Normalized & Filtered, Target-Agnostic Features Only**
1. Movement time divided by the Index of Difficulty	1. Movement time divided by the Index of Difficulty (normalized)	1. Movement time divided by movement distance	1. Movement time divided by movement distance (normalized)
2. Movement time divided by logarithm of movement distance	2. Movement time divided by logarithm of movement distance (normalized)	2. Movement error	2. Movement time divided by logarithm of movement distance (normalized)
3. Movement error	3. Movement error	3. Relative time to the first peak in speed	3. Movement error
4. Movement variability	4. Movement variability	4. Relative time to the first peak in acceleration	4. Movement variability
5. Relative time to the first peak in speed	5. Number of submovements	5. Relative time to the maximum peak in acceleration	5. Number of submovements (normalized)
6. Relative time to the maximum peak in speed	6. Number of submovements (normalized)	6. Relative time to the maximum peak in jerk	6. Relative time to the first peak in speed
7. Relative time to the first peak in acceleration	7. Relative time to the first peak in acceleration		7. Relative time to the first peak in acceleration
8. Relative time to the maximum peak in jerk			8. Relative time to the maximum peak in acceleration
			9. Relative time to the maximum peak in jerk

Table 3. Sets of features that resulted in best performance in the four approaches evaluated.

The result of the feature selection process was a classifier using 8 out of 34 available features (shown in Table 3). Two of the features relate to the speed of movement execution, two are related to the shape of the path, and four use information captured in the speed, acceleration, and jerk profiles.

On average, 25% of each participant's natural movements were classified as deliberate. The errors in estimating MT/ID and the Fitts' model predictions were reduced by factors of 8 and 7, respectively, to 22%. This improvement translated into more reliable results of statistical tests: the probability of declaring a statistically significant difference where none was present was reduced nearly by half and there were no cases where a statistically significant result was declared in the direction opposite to what was observed in the experimental data. The probability of missing a statistically significant result was reduced by a third, but still remained high.

The overall recall was 89%, meaning that on average 11% of each participant's experimental movements were classified as distracted. Because the data collection occurred in an unsupervised setting, this number may reflect the true quality of our experimental data. What is a source of concern is the fact that for one of the participants the recall was only 25%. This was the only such extreme outlier—the next lowest number was 69%. An inspection of the data revealed that the participant with the lowest recall was P44, who was the oldest person to take part in our study and who reported mechanical difficulties with her mouse. This participant was by far the slowest of the 18 as measured by the Movement Time divided by the Index of Difficulty. P44 was, unsurprisingly, also the outlier in terms of the fraction of natural movements classified as deliberate: only 4% of her movements were classified as such.

Approach 2: Mitigating the Impact of Individual Differences

The results presented above suggest that the approach, as presented so far, is sensitive to individual differences among participants: it identifies the best exemplars of deliberate movements across the population. Instead, in most situations the desired behavior would be to identify those movements that reflected a particular individual's best attempt at deliberate, targeted pointing movements.

To mitigate the effects of such individual differences, we propose a second approach where in addition to the movement trajectory and the size of the target, we also have available a set of trajectories representative of that user's typical natural computer behavior.

In this approach, we normalized each participants' data by subtracting (separately for each feature) the mean value of the feature over natural data and then dividing the resulting value by the standard deviation (also computed over the natural data). This normalization procedure eliminated participant-specific differences in magnitude while preserving the relative differences between natural and experimental data. We then created a new set of data that included both the original features and the per-participant normalized counterparts for a total of 68 features per movement.

We then re-run the feature selection procedure. The recall results for the resulting classifier indicate that the effects of differences in individual motor performance have been indeed attenuated: the recall is still the lowest for P44, but it is now at 65% — a value that may reflect the participant's problems with a "sticky" mouse more than the limitations of the algorithm. The fraction of P44's natural movements classified as deliberate is now at 27%.

Compared to the initial filtering method proposed, the normalization procedure improved both the averages and the outlier results for all the metrics. This provides further evidence that normalization improves the generalizability of the approach making it more robust to the individual differences in motor performance. The normalization procedure does, however, implicitly rely on there being a similar distribution of deliberate and distracted movements in each user's natural interactions with the computer.

Approaches 3 & 4: Filtering with Target-Agnostic Features

Because most operating systems make it easy to capture all pointer-related events, but reveal incomplete information about the presence, location, and size of interactors [2, 5, 9], in some cases researchers may not have access to the information about the locations and sizes of targets, but may still want to collect instances of lab-quality pointing movements. The earlier two approaches both made use of the information about the size of the target: Movement Time divided by

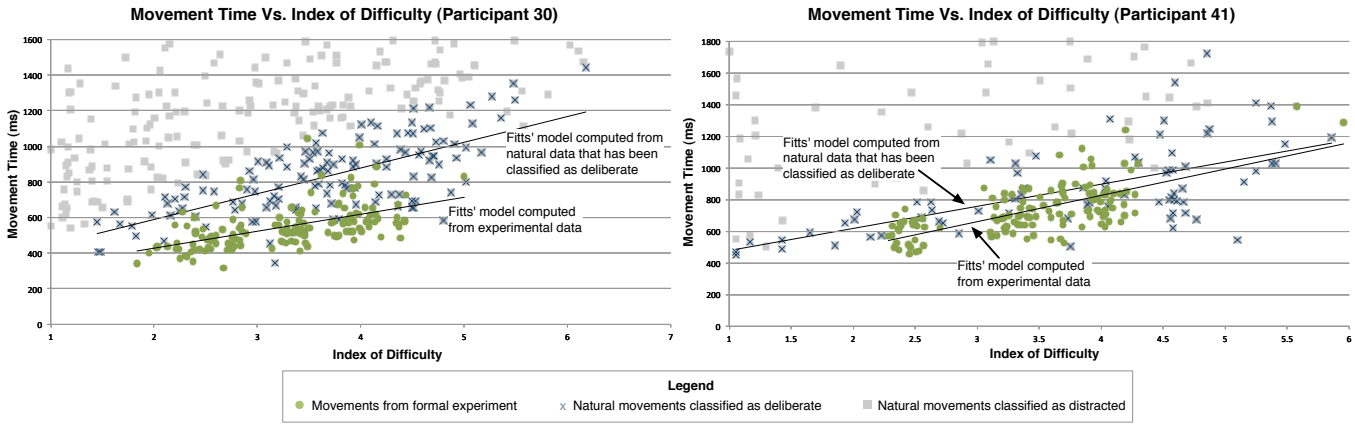

Figure 2. Understanding errors

the Index of Difficulty was a feature included in both models. Could we determine if a movement was deliberate in a target-agnostic manner?

To answer this question, we repeated our analysis using only target-agnostic features. We used the same evaluation criteria as in the previous analyses: although we recognize that target-agnostic classification will not be used for building Fitts' law models, these measures indirectly indicate whether deliberate, targeted movements are being identified correctly.

The result are two new approaches: Approach 3, which can perform classification based just on the movement trajectory information, and Approach 4, which (like Approach 2) requires a sample of trajectories representative of the user's natural interactions with the computer.

The features selected for the two resulting approaches are listed in Table 3 and the results are summarized in the last two columns of Table 2. Unsurprisingly, these approaches in general do not perform as well as the previous two. In particular, they result in a high rate of false positives in the tests for statistically significant differences. We have less confidence that these two approaches can be used to reliably filter individual interactions. However, we find that 86% of classification decisions of Approach 3, and 93% of Approach 4, agree with the predictions of Approach 2.

Understanding Sources of Error

Lastly, we take another close look at the results to understand the causes of the remaining errors — do they point to any specific limitations of our methods? To gain some insights, we review the data of Participant 30 for whom both Approach 1 an 2 resulted in consistently high errors in the estimates of MT/ID and Index of Performance — Participant 30 was not an extreme outlier like Participant 44 discussed earlier, but the error rates computed from her data were high and did not improve from Approach 1 to Approach 2. The left panel in Figure 2 contrasts the experimental data from this participant with her natural data classified as deliberate using Approach 2. What this figure makes apparent, is that Participant 30 rarely moved as quickly in a natural setting as she did during the experiment. A follow-up revealed that, when performing the formal experimental tasks, this participant "raced" a friend who had also signed up for our study. Thus, the apparent errors in our data can likely be explained by her extreme effort level during the formal experiment.

Our approach cannot compensate for such situations. In our data set, however, this was an unusual situation and the data from our other participants tended to look more like that of Participant 41 (illustrated the right panel of Figure 2), where the deliberate movements in the natural setting were similar to the movements observed in the formal experiment.

CONCLUSIONS AND FUTURE DIRECTIONS

We have developed and evaluated four user-independent classifiers (each making different assumptions about the type and quantity of data available) capable of discriminating between deliberate, targeted mouse pointer movements and those that were performed when the participant was in any way distracted.

Our results demonstrate that the classifiers we have developed can be used to filter pointing data collected in naturalistic settings such that the resulting data have properties very similar to the data typically collected in lab settings. This enables some of the measurements that are currently performed in laboratory settings to be conducted in situ, allowing for larger-scale and longer-term observations.

In accessibility-related settings, in situ pointing performance modeling can be used to adapt user interfaces to people's changing abilities [7, 6] and to understand how people's abilities change over time on different time scales (daily due to medication or fatigue and over longer periods of time due to progression of the disease or the effects of therapy) [11]. However, as the results for Approach 1 have demonstrated, our methodology will need further refinement before it can generalize to people with substantially atypical motor abilities.

The methods we have described here provide a blueprint that can be used to develop similar classifiers for other interactions

(e.g., dragging, steering) and for different types of hardware (e.g., touch screens).

Lastly, the methods we have presented may be used to assess the quality of the data obtained from on-line experiments conducted with unsupervised remote participants: for experiments involving mechanical pointing tasks (i.e., tasks where participants are not simultaneously solving a complex cognitive task or need to engage in visual search), our classifier can be used used to identify and exclude from the analysis participants who appear to have been distracted during the experiment.

ONLINE APPENDIX

To enable others to use, validate and extend our results, we also make the following resources available from our web site at http://iis.seas.harvard.edu/resources:

- the user-independent classifiers already trained on more than 2000 deliberate movements and several thousands distracted movements collected in realistic conditions from 18 diverse participants;
- the data set containing 2762 experimental and 5085 natural movement trajectories;
- and the source code of the software used in this work.

Acknowledgments. Evan Greif led the implementation of the study apparatus. We thank Dimitrios Antos, Kanya (Pao) Siangliulue, Kenneth Arnold, Jennifer Mankoff, Tovi Grossman, and the anonymous CHI reviewers for helpful comments on earlier drafts of this manuscript. Katharina Reinecke was supported by the Swiss National Science Foundation under fellowship number PBZHP2-135922.

REFERENCES

1. Accot, J., and Zhai, S. Refining Fitts' law models for bivariate pointing. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press (New York, NY, USA, 2003), 193–200.
2. Chapuis, O., Blanch, R., and Beaudouin-Lafon, M. Fitts' law in the wild: A field study of aimed movements. Tech. Rep. 1480, LRI, Univ. Paris-Sud, France, December 2007. 11 pages.
3. Denis, F., Gilleron, R., and Letouzey, F. Learning from positive and unlabeled examples. *Theor. Comput. Sci. 348*, 1 (2005), 70–83.
4. Elkan, C., and Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, ACM (New York, NY, USA, 2008), 213–220.
5. Evans, A., and Wobbrock, J. Taming wild behavior: The Input Observer for obtaining text entry and mouse pointing measures from everyday computer use. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '12). Austin, Texas (May 5-10, 2012)*, ACM Press (2012). To appear.
6. Gajos, K. Z., Weld, D. S., and Wobbrock, J. O. Automatically generating personalized user interfaces with Supple. *Artificial Intelligence 174* (2010), 910–950.
7. Gajos, K. Z., Wobbrock, J. O., and Weld, D. S. Improving the performance of motor-impaired users with automatically-generated, ability-based interfaces. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ACM (New York, NY, USA, 2008), 1257–1266.
8. Grossman, T., and Balakrishnan, R. A probabilistic approach to modeling two-dimensional pointing. *ACM Transactions on Computer-Human Interaction (TOCHI) 12*, 3 (2005), 435–459.
9. Hurst, A., Hudson, S. E., and Mankoff, J. Automatically identifying targets users interact with during real world tasks. In *Proceedings of the 15th international conference on Intelligent user interfaces*, IUI '10, ACM (New York, NY, USA, 2010), 11–20.
10. Hurst, A., Hudson, S. E., Mankoff, J., and Trewin, S. Automatically detecting pointing performance. In *IUI '08: Proceedings of the 13th international conference on Intelligent user interfaces*, ACM Press (Jan. 2008).
11. Hurst, A., Mankoff, J., and Hudson, S. E. Understanding pointing problems in real world computing environments. In *Assets '08: Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, ACM (New York, NY, USA, 2008), 43–50.
12. International Organization for Standardization. 9241-9 Ergonomic requirements for office work with visual display terminals (VDTs)-Part 9: Requirements for non-keyboard input devices, 2000.
13. Jagacinski, R., Repperger, D., Moran, M., Ward, S., and Glass, B. Fitts' law and the microstructure of rapid discrete movements. *Journal of Experimental Psychology: Human Perception and Performance 6*, 2 (1980), 309–320.
14. Kaiser, J. F., and Reed, W. A. Data smoothing using low-pass digital filters. *Review of Scientific Instruments 48*, 11 (1977), 1447–1457.
15. MacKenzie, I. S. Fitts' law as a research and design tool in human-computer interaction. *Human-Computer Interaction 7*, 1 (1992), 91–139.
16. MacKenzie, I. S., Kauppinen, T., and Silfverberg, M. Accuracy measures for evaluating computer pointing devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '01, ACM (New York, NY, USA, 2001), 9–16.
17. MacKenzie, S. I., and Buxton, W. Extending Fitts' law to two-dimensional tasks. In *CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press (New York, NY, USA, 1992), 219–226.
18. Meyer, D. E., Abrams, R. A., Kornblum, S., Wright, C. E., and et al. Optimality in human motor performance: Ideal control of rapid aimed movements. *Psychological Review 95*, 3 (1988), 340–370.
19. Rzeszotarski, J. M., and Kittur, A. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, ACM (New York, NY, USA, 2011).
20. Walker, N., Philbin, D. A., and Fisk, A. D. Age-related differences in movement control: adjusting submovement structure to optimize performance. *J Gerontol B Psychol Sci Soc Sci 52*, 1 (January 1997).
21. Wobbrock, J. O., and Gajos, K. Z. Goal crossing with mice and trackballs for people with motor impairments: Performance, submovements, and design directions. *ACM Trans. Access. Comput. 1*, 1 (May 2008), 1–37.
22. World Wide Web Consortium. Accessible Rich Internet Applications (WAI-ARIA) 1.0, 2011. Accessed on January 8, 2012.
23. Yu, H., Han, J., and Chang, K. C.-C. Pebl: positive example based learning for web page classification using svm. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM (New York, NY, USA, 2002), 239–248.