Gaze-Based Proactive User Interface for Pen-Based Systems

Çağla Çığ
Koç University
Istanbul, Turkey
ccig@ku.edu.tr

ABSTRACT
In typical human-computer interaction, users convey their intentions through traditional input devices (e.g. keyboards, mice, joysticks) coupled with standard graphical user interface elements. Recently, pen-based interaction has emerged as a more intuitive alternative to these traditional means. However, existing pen-based systems are limited by the fact that they rely heavily on auxiliary mode switching mechanisms during interaction (e.g. hard or soft modifier keys, buttons, menus). In this paper, I describe the roadmap for my PhD research which aims at using eye gaze movements that naturally occur during pen-based interaction to reduce dependency on explicit mode selection mechanisms in pen-based systems.

Author Keywords: Sketch-based interaction; multimodal interaction; predictive interfaces; gaze-based interfaces; feature representation; multimodal databases

ACM Classification Keywords
H.1.2 **[Models and Principles]**: User/Machine Systems – *Human information processing;* H.5.2 **[Information Interfaces and Presentation]**: *Input devices and strategies, Interaction styles, User-centered design*

INTRODUCTION
People commonly prefer pen and paper for brainstorming, for exchanging ideas with others, or simply for taking notes. Due to this tendency, pen-based user interfaces promise a more intuitive and accessible alternative to traditional graphical user interfaces. Via pen-based interfaces, users can virtually *produce* and *manipulate* various kinds of free-form sketches (e.g. flowcharts, family trees, and electrical circuit diagrams). Some examples to frequently employed virtual manipulation tasks are dragging, maximizing, or minimizing individual sketch parts or scrolling the whole sketch canvas. During a typical interaction scenario, users repeatedly alternate between sketching and these manipulation tasks. However, prior to sketching or performing a manipulation task, users need to specify the intended mode of interaction via various auxiliary mode switching mechanisms (e.g. multi-finger gestures, context/pop-up menus, and external buttons). For example, in pen-enabled smart phones, users are forced to put the pen aside and switch to multi-finger gestures for many tasks (e.g. spread/pinch for zoom in/out and swipe to navigate back/forward). These gestures require the simultaneous use of 2, 3, or even 4 fingers [1] (Figure 1). The necessity of switching between pen and multi-touch input goes against the goal of seamless interaction in pen-based devices.

In this paper, I describe my PhD research plan and progress which is centered around a novel multimodal approach to alleviate dependence on explicit mode switching in pen-based systems. The first part of my PhD research explores whether gaze movements that naturally accompany pen-based user interaction can be used to infer a user's task-related intentions and goals. Based on our preliminary results indicating a connection between gaze movements and pen-based interaction tasks, for the second part of my PhD research, we envision a proactive system capable of actively monitoring user's eye gaze and pen input to detect the intention to switch modes in an online setting, and act accordingly.

Figure 1. Switching between pen and multi-touch input for object manipulation (e.g. image resizing) in pen-enabled smart phones.

RESEARCH OVERVIEW
Our overall approach to gaze-based proactive user interfaces for pen-based systems consists of two parts: offline part where we build our gaze-based virtual task prediction system and online part where we integrate our prediction system to build a proactive user interface. Both parts entail interpreting pen and eye gaze input within a machine learning framework.

ICMI '14, November 12 - 16, 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2663204.2666287

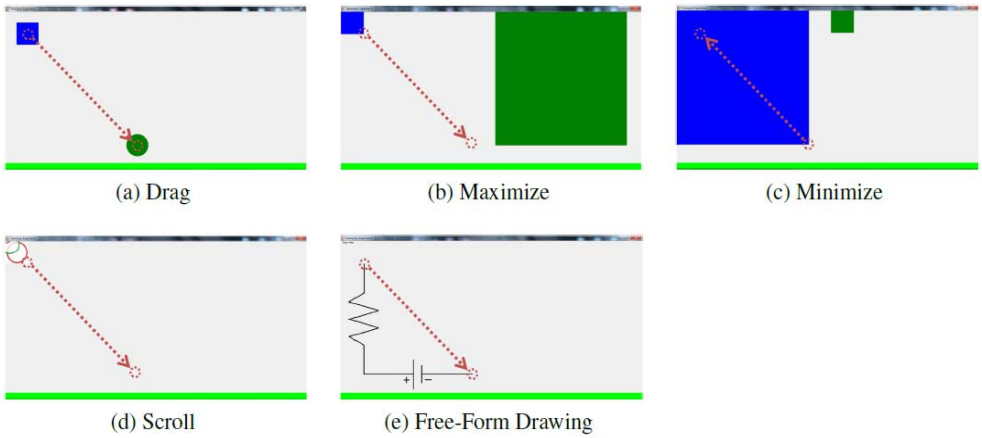


Figure 2. Pen-based virtual interaction tasks included in our research. Starting and ending regions of desired pen motion in each task are visualized with dotted circles. In the rest of the article, the center points of these regions will be referred to as *anchor points*. Direction of the desired pen motion in each task is visualized with a dotted arrow connecting the starting and ending regions. It is important to note that the dotted visualizations only serve as a reference within this paper and they are not shown to the user during data collection.

This primarily requires large amounts of data for training classifiers. Therefore, initially we have collected sketch and gaze data during a number of pen-based interaction tasks and built a multimodal database. Detailed description and discussion of each part can be found in the following sections.

Multimodal Data Collection (Completed)

We have collected data in a controlled setup where the users were asked to carry out a number of frequently employed pen-based virtual interaction tasks (Figure 2). To create a database composed of synchronized sketch and gaze data, we used a tablet and a Tobii X120 stand-alone eye tracker for the sketch and gaze modalities, respectively. Multimodal data was collected across three different scales to test our system in terms of scale-invariance. The scale variable determines the length of the path connecting the two anchor points present in each task (Figure 2). In light of facts about human vision, lengths of the paths were set to 21 cm, 10.5 cm, and 5.25 cm for the *large*, *medium*, and *small* scales, respectively. We refer to each run of a certain task at a certain scale as *a task instance*. Our multimodal database consists of 1500 task instances collected from 10 participants (6 males, 4 females) over 10 randomized repeats of 5 tasks across 3 scales. This carefully compiled database is the first of its kind, and we believe it will serve as a reference database for future research on the topic.

Offline Part (Ongoing)

Briefly, our task prediction system will be built as follows: We will extract novel gaze-based features from our multimodal database and train a task prediction model using supervised machine learning techniques. These steps will be executed only once. Then, our system will be ready for online prediction.

Novel Gaze-Based Feature Representation

Our system will utilize only two kinds of features for gaze-based task prediction: *Instantaneous Distance Between Sketch and Gaze Positions* and *Within-Cluster Variance of Gaze Positions*. The strength of these features stems from the fact that they eliminate the need for possibly subject- and interface-specific preprocessing steps common in gaze-based systems. Some examples of these common error-prone preprocessing steps include segmentation of gaze data into fixations and saccades and manual specification of regions of interest.

Feature 1: Instantaneous Distance Between Sketch and Gaze Positions

Hand-eye coordination behavior inherent in virtual interaction tasks changes over the course of a task instance as a function of changes in user sub-tasks [2, 3, 4, 5]. The multiple steps of each task can be thought of as consecutive sub-tasks and each sub-task entails a different type of hand-eye coordination behavior. The rationale behind the first feature of our novel gaze-based feature representation is based on this observation and attempts to capture the goal-dependent dynamic aspects of human hand-eye coordination behavior through the evolution of the distance between instantaneous gaze and sketch positions calculated over a task instance.

Consider the task in Figure 2a. In a typical instance of this task, the user is instructed to drag a source object (the blue square) onto a target object (the green circle). The sub-tasks of this task are 1) positioning the pen on the source object, 2) determining the position of the target object and 3) dragging the source object towards the target object. We argue that the dynamic aspects of human hand-eye coordination behavior are reflected in the distance values between instantaneous gaze and sketch positions calculated over time. Figure 3 and Figure 4 generated from the same sample task instance support our argument. Figure 3 gives a visualization of the user's sketch data along with a number of sketch and gaze data samples. Sketch and gaze data points collected at identical time instances are connected with dotted lines. The length of a connection line denotes

the value of the sketch-gaze distance feature for the corresponding instance. Figure 4 demonstrates how the value of this feature changes over time. In this figure, the sketch-gaze distance feature is plotted for the same user and same task, over three different scales. Peaks of the plots could conceivably mark the second sub-task during which the user, after having positioned the pen on the source object, is now gazing at the target object. Note that sketch-gaze distance feature expresses similar characteristics across different scales; thus our approach and our novel feature can be generalized and applied to data collected across different scales.

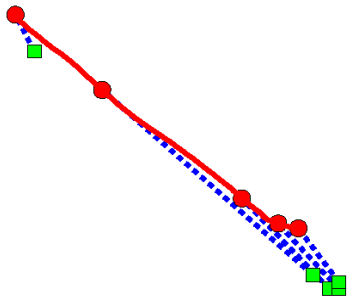


Figure 3. Visualization of the user's sketch data (solid line) along with a number of sketch (circles) and gaze (squares) data samples. Dotted lines connect the instantaneous sketch and gaze sample pairs.

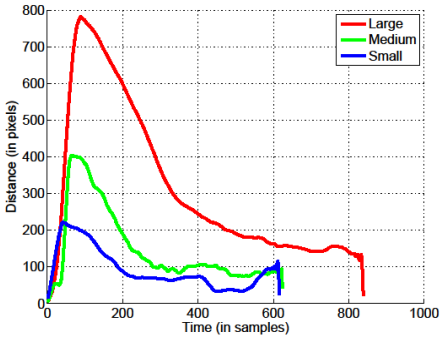


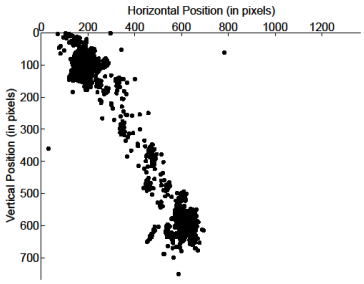
Figure 4. Visualization of the changes in the value of sketch-gaze distance feature as a function of time.

Inspection of the sketch-gaze distance curves for the *drag* task reveals that the rapid rise and gradual decline behavior is typical of all instances of the *drag* task. Similarly, the distance curves for the other tasks also display task-specific characteristic rise and fall behaviors. We will compute distinct sketch-gaze distance curves for each virtual interaction task using sketch-gaze distance curves of all task instances.

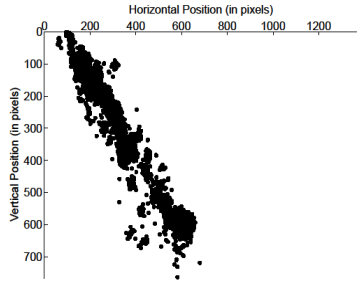
Feature 2: Within-Cluster Variance of Gaze Positions

Humans employ two different modes of voluntary gaze shifting mechanism to orient the visual axis. These modes are referred to as saccadic and smooth pursuit eye movements. It is widely accepted that "saccades are primarily directed toward stationary targets whereas smooth pursuit is elicited to track moving targets" [6]. Typical virtual interaction tasks contain both stationary and moving targets. A user's attention can be dominantly directed towards targets of either type depending on the intended task.

Our experiments show that in a typical *drag* task, saccades are more common and the user's attention is drawn from one stationary target which is the initial position of the object currently being dragged to the other stationary target which is the intended position of the object (Figure 5a). Conversely during *free-form drawing* (Figure 5b), smooth pursuit is more common and the user's attention is drawn to the moving target (the newly appearing ink). In saccades, gaze points accumulate around the stationary targets whereas in smooth pursuit, gaze points scatter along the pursuit path. The second feature of our novel gaze-based feature representation is based on these observations, and hence attempts to quantify how the data is structured in terms of saccades and fixations.



(a) Gaze data for the *drag* task. Saccadic eye movements result in gaze point clusters with low within-cluster variance.



(b) Gaze data for the *free-form drawing* task. Smooth pursuit eye movements result in gaze point clusters with high within-cluster variance.

Figure 5. Gaze data corresponding to 10 repeated task instances of a user.

Intention Prediction and Evaluation

We plan to conduct comprehensive tests to evaluate the effectiveness of the features introduced above in predicting virtual interaction tasks. During evaluation, we will focus on several aspects, including the prediction accuracy and scale-invariance. In addition, we will run feature selection tests to evaluate the relevance and redundancy of the features introduced above.

Practical usage of our prediction system may involve a range of display devices and user interfaces with varying sizes and constraints. Robustness of a feature representation to scale variances is important if we want our prediction

system to work equally accurately despite these variances. Therefore, we plan to conduct *scale-invariance tests* to evaluate the robustness of our feature representation to such variances in scale.

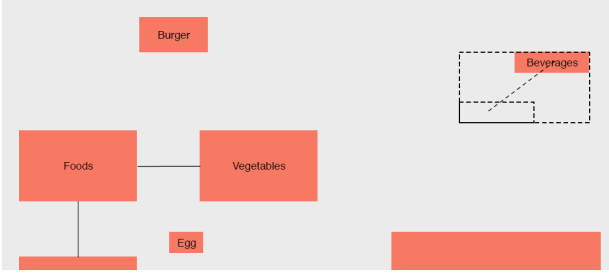
We will also compare the prediction power of our novel gaze-based feature representation to that of commonly utilized and well-established sketch-based feature representations in the literature, namely IDM Features [7] and Zernike Moments [8]. Lastly, we will explore whether gaze-based and sketch-based feature representations can be combined via classifier-level or fusion-level fusion methods to increase prediction accuracy.

Online Part (Ongoing)

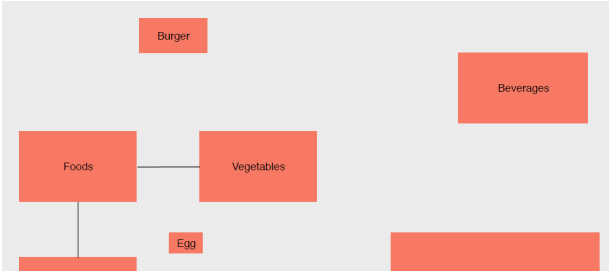
In the light of preliminary results indicating a connection between gaze movements and pen-based interaction tasks, the second part of our research explores the feasibility of using our prediction system to build a proactive user interface. When the user performs a pen action (demarcated by a pen-down and a pen-up event), the planned proactive user interface will actively detect and switch to the currently intended mode of interaction based on user's synchronized pen trajectory and eye gaze information during pen-based interaction. Intention predictions will be carried out by the previously trained model and the features extracted from the corresponding sketch-gaze data of the user.

The biggest challenge we face here is concerned with providing feedback. In line with the feedback principle of design [9], while the user is performing a pen action, the user interface must provide immediate and appropriate visual feedback about the effects of user's actions and do this without causing any changes in user's natural eye gaze behavior. However, the effects of user's actions depend on user's task-related intentions and goals, which are not known to the interface until the action is completed. Therefore, the interface must provide feedback about user's intentions, from the start to the end of a pen action, without knowing user's intentions. To this end, we propose a proactive user interface approach where effects of all possible actions are visualized simultaneously for the duration of an action. Variety of possible actions at any instant depends on the context and these context-dependent action rules will be defined in the later phases of our research. When the action is finalized, irrelevant effects disappear and only the effects of the predicted action remain visible.

To realize this approach, we introduce a practical application scenario that involves frequent use of all five tasks currently distinguishable by our task prediction system. This scenario that we refer to as *The Categorical Tree Application* is a type of diagramming application that can be used for mind mapping purposes (Figure 6). The application interface consists of rectangles with arbitrary dimensions. These rectangles represent conceptual categories, sub-categories, and sample objects presumably created quickly by the user at the beginning of the mind mapping process. The user's task at this point is to organize the mind map by dragging and resizing rectangles, connecting them as appropriate, and scrolling the application interface if necessary. The user is free to perform these sub-tasks naturally in any desired order, and without any constraints. More importantly, the user does not have to make a specific gesture or locate the correct button to repeatedly switch the interaction mode in-between these sub-tasks.



(a) The interface during the pen action on *Beverages* category. Effects of all possible actions (i.e. *drag* and *maximize*) are visualized simultaneously with dotted lines.



(b) The interface at the end of the pen action visualizing only the effects of the predicted action (i.e. *maximize*).

Figure 6. The Categorical Tree Application. The user is able to issue a *maximize* command without going into the trouble of switching modes beforehand.

CONTRIBUTIONS AND FUTURE WORK

We have proposed a novel multimodal approach to alleviate dependence on explicit mode switching in pen-based systems. This approach initially entails building a gaze-based virtual task prediction system that infers intended user actions by monitoring and analyzing eye gaze movements that users naturally exhibit during pen-based user interaction. Our prediction system will open the way for more natural user interface paradigms where the role of the computer in supporting interaction is to "interpret user actions and [do] what it deems appropriate" [10]. It is widely accepted that intelligent mode selection mechanisms that provide low cost access to different interface operations will dominate new user interface paradigms [11].

Our first contribution is a carefully compiled multimodal dataset that consists of eye gaze and pen input collected from participants completing various virtual interaction tasks. Our second contribution is a novel gaze-based feature representation, which is rooted in our understanding of

human perception and gaze behavior. Our feature representation is neither subject- nor interface-specific, and is expected to perform better than common, well-established sketch recognition feature representations in the literature. Our third contribution is expected to be a novel gaze-based task prediction system based on this feature representation that can generalize to variations in task type and scale.

Further experiments will be required to evaluate the usability aspects of our proactive user interface, and compare it to the state of the art mode switching mechanisms in the literature. However, there are a number of major issues we need to address beforehand. First, we need to find a way to handle prediction errors. Although our intention prediction system is expected to be fairly accurate (with a success rate of more than 80%), inaccurate predictions will still be possible. Therefore, further research is required to investigate approaches for detecting and recovering from system errors. Otherwise, users might confuse system errors with user-induced errors and diverge from natural gaze behavior in an effort to avoid them. In turn, this divergence will conceivably reduce the quality of the user's experience with the interface as well as the accuracy of our prediction system that assumes natural user behavior. Second issue we need to address concerns visualization. We envision a proactive user interface where effects of all possible actions are visualized simultaneously until a pen action is finalized and a prediction is made. However, showing the effects of irrelevant actions for the entire duration of a pen action can be cumbersome and lead to a heavily cluttered interface as the number of possible actions increases. In consequence, several questions remain to be addressed with respect to visualization of user's task-related intentions and goals: Can we benefit from eager recognition techniques to avoid waiting until the end of a pen action to make a prediction? Can we use increasing levels of transparency to indicate decreasing likelihoods of a pen action being the intended pen action, i.e. highly probable actions become more emphasized as unlikely actions fade out? Formal user studies will be needed to obtain definitive answers to such questions. Lastly, this research was concerned with pen-based systems; however, the results should be applicable also to other kinds of pointer-based systems that accept stylus, finger, or mouse input. However, more experiments will be needed to verify whether our task prediction system or a similar system inspired by our current findings generalizes well to such systems.

ACKNOWLEDGMENTS

The author would like to thank Dr. T. Metin Sezgin for his invaluable guidance and feedback. The author appreciates the assistance of Eren Sezener regarding the design of the Categorical Tree Application. Financial supports from TÜBİTAK (The Scientific and Technological Research Council of Turkey) under grant number 110E175 and TÜBA (Turkish Academy of Sciences) are gratefully acknowledged.

REFERENCES

1. Samsung Galaxy Note 2 [Online image]. Retrieved September 3, 2014, from http://www.cnet.com/products/samsung-galaxy-note-2/2.
2. Hayhoe, M. and Ballard, D. Eye movements in natural behavior. *Trends in Cognitive Sciences 9*, 4 (2005), 188-194.
3. Fathi, A., Li, Y. and Rehg, J.M. Learning to recognize daily actions using gaze. In *Proc. ECCV 2012*, Springer-Verlag (2012), 314-327.
4. Johansson, R.S., Westling, G., Bäckström, A. and Flanagan, J. R. Eye-hand coordination in object manipulation. *The Journal of Neuroscience 21*, 17 (2001), 6917-6932.
5. Ballard, D.H., Hayhoe, M.M., Li, F., Whitehead, S.D., Frisby, J.P., Taylor, J.G. and Fisher, R.B. Hand-eye coordination during sequential tasks [and discussion]. *Philosophical Transactions: Biological Sciences 337*, 1281 (1992), 331-339.
6. Orban de Xivry, J.-J. and Lefèvre, P. Saccades and pursuit: two outcomes of a single sensorimotor process. *The Journal of Physiology 584*, 1 (2007), 11-23.
7. Ouyang, T.Y. and Davis, R. A visual approach to sketched symbol recognition. In *Proc. IJCAI 2009*, Morgan Kaufmann Publishers Inc. (2009), 1463-1468.
8. Khotanzad, A. and Hong, Y.H. Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*, 5 (1990), 489-497.
9. Norman, D.A. *The design of everyday things.* Basic Books, 2002.
10. Nielsen, J. Noncommand user interfaces. *Communications of the ACM 36*, 4 (1993), 83-99.
11. Negulescu, M., Ruiz, J. and Lank, E. Exploring usability and learnability of mode inferencing in pen/tablet interfaces. In *Proc. SBIM 2010*, Eurographics Association (2010), 87-94.