# Gaze-Based Virtual Task Predictor

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

Tevfik Metin Sezgin
Koç University
Istanbul, Turkey
mtsezgin@ku.edu.tr

## ABSTRACT
Pen-based systems promise an intuitive and natural interaction paradigm for tablet PCs and stylus-enabled phones. However, typical pen-based interfaces require users to switch modes frequently in order to complete ordinary tasks. Mode switching is usually achieved through hard or soft modifier keys, buttons, and soft-menus. Frequent invocation of these auxiliary mode switching elements goes against the goal of intuitive, fluid, and natural interaction. In this paper, we present a gaze-based virtual task prediction system that has the potential to alleviate dependence on explicit mode switching in pen-based systems. In particular, we show that a range of virtual manipulation commands, that would otherwise require auxiliary mode switching elements, can be issued with an 80% success rate with the aid of users' natural eye gaze behavior during pen-only interaction.

## Categories and Subject Descriptors
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*

## General Terms
Algorithms, Human Factors.

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

## 1. INTRODUCTION
People commonly prefer pen and paper for brainstorming, for exchanging ideas with others, or simply for taking notes. This makes pen-based user interfaces a promising, more intuitive and accessible alternative to traditional graphical user interfaces. Using pen-based interfaces, users can *produce* and *manipulate* various kinds of free-form sketches (e.g. flowcharts, family trees, and electrical circuit diagrams). In the rest of the paper, commands issued by users to manipulate virtual objects during pen-based interaction will be referred to as *virtual manipulation tasks*. 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). Even high-end graphics tablets preferred mainly by digital artists such as Wacom Cintiq 24HD [1] (Figure 1) lack purely pen-based interaction. For many tasks (e.g. scroll, zoom in/out, navigate back/forward), users are forced to use on-pen or on-tablet external buttons called "express keys", "touch rings", and "radial menus". This requirement of switching back and forth between pen and button-based input interrupts the interaction flow, and hence causes an overall disappointing experience.

**Figure 1. On-pen or on-tablet external buttons in pen-enabled graphics tablets serve as an example to pen-based devices with interaction paradigms that are not purely pen-based.**

In this paper, we present a novel multimodal approach to reduce reliance on explicit mode switching mechanisms in pen-based systems. Our current findings demonstrate that we can use gaze movements that naturally accompany pen-based user interaction to predict a user's task-related intentions and goals. Based on this connection between gaze movements and pen-based interaction tasks, 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. We believe that the non-intrusive and transparent use of gaze modality for intention prediction alleviates dependence on explicit mode switching and takes us one step towards the goal of natural pen-based interaction. There is no existing piece of work that uses eye gaze information as we do, thus the presented research and our approach is highly novel.


*GazeIn'14*, November 16, 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2666642.2666647

Briefly, our task prediction system is built as follows: Initially we collect sketch and gaze data during a number of pen-based interaction tasks and build a multimodal database (Section 2). We then extract novel gaze-based features from this database (Section 3) and train a task prediction model using supervised machine learning techniques (Section 4). These steps are executed only once. Then, our system is ready for online prediction.

## 2. MULTIMODAL DATA COLLECTION

Our task prediction system interprets sketch and gaze input within a supervised machine learning framework. This primarily necessitates compiling a large database for training classifiers. For this purpose, we ask users to carry out a number of frequently employed pen-based virtual interaction tasks in a controlled setup (Table 1, left column). We give the following instructions to the users for each task:

- *Drag:* Drag the blue square onto the center of the green circle.
- *Maximize:* Increase the size of the blue square to match the size of the green square.
- *Minimize:* Decrease the size of the blue square to match the size of the green square.
- *Scroll:* Pull the chain until the color of the last link is clearly visible.
- *Free-form drawing:* Connect the battery and the resistor with a wire.

Our physical setup for collecting synchronized sketch and gaze data consists of a tablet and a Tobii X120 stand-alone eye tracker for the sketch and gaze modalities, respectively (Figure 2). We collected sketch and gaze data across three different scales, where the scale variable determines the length of the task trajectory. In light of facts about human vision, we set the lengths of the trajectories to 21 cm, 10.5 cm, and 5.25 cm for the *large*, *medium*, and *small* scales, respectively. In the rest of the paper, 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.

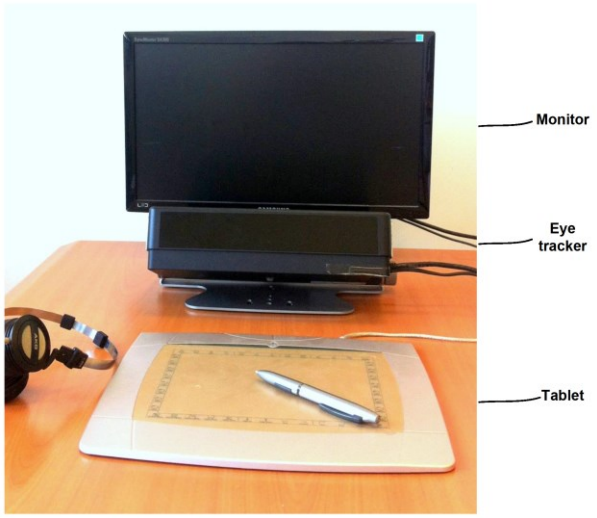


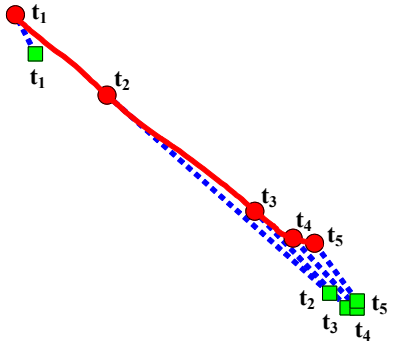

**Figure 2. Physical setup used for multimodal data collection.**

## 3. NOVEL GAZE-BASED FEATURE REPRESENTATION

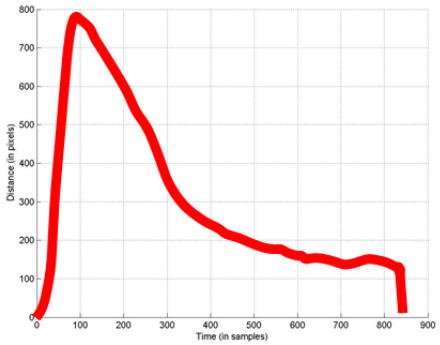
Our framework employs just two kinds of features to predict pen-based virtual interaction tasks using gaze information: *Instantaneous Distance Between Sketch and Gaze Positions* and *Within-Cluster Variance of Gaze Positions*. The advantage of our feature representation over related gaze-based feature representations lies in the fact that our features eradicate the need for possibly subject- and interface-specific preprocessing steps. Some of the commonly utilized error-prone preprocessing steps can be listed as segmentation of gaze data into fixations and saccades and manual specification of regions of interest.

### 3.1 Feature 1: Instantaneous Distance Between Sketch and Gaze Positions

Let $G_t < x, y >$ represent the 2D position of the gaze on the screen at time $t$ during the execution of a particular task (Figure 3a). Similarly, let $P_t < x, y >$ denote the 2D position of the stylus tip on the drawing device at time $t$ (Figure 3a). Finally, let $D_t = |G_t - P_t|$ be the distance between these points (Figure 3b). We argue that throughout the completion of a task instance, $D_t$ evolves in a strongly task-dependent fashion. More specifically, if we compute the distance curves $D_t$ for all task instances of a certain task type, we see that these curves have similar rise/fall characteristics. On the other hand, for different task types, the profiles of distance curves are quite different from each other.




**a) Visualization of the user's sketch data (solid line) along with a number of $G_t < x, y >$ samples (squares) and a number of $P_t < x, y >$ samples (circles). In addition, instantaneous gaze and sketch samples are connected with dotted lines.**

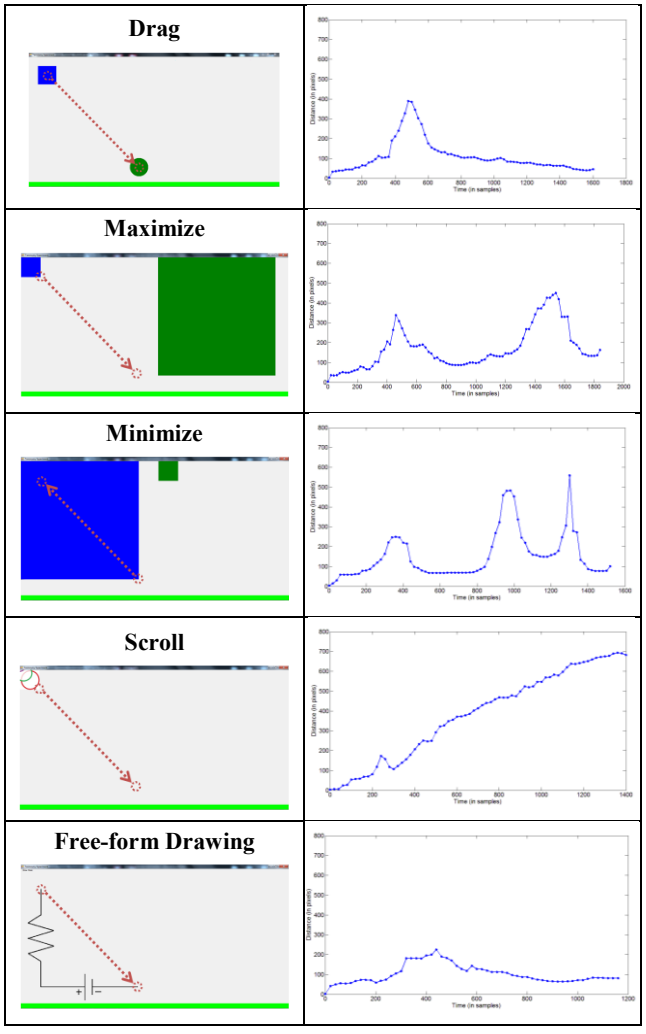



**b) Evolution of the distance curve $D_t$ over time.**

**Figure 3. Plots visualizing the computation of the distance curve for a given task instance.**

However, even for the same task type, the rates at which distance curves evolve are different; therefore these curves are not completely identical. In order to solve this problem, we need to align sketch-gaze distance curves that have similar rise/fall characteristics but evolve at different rates. To this end, we use dynamic time warping [2] and acquire a characteristic curve for each task type (Table 1, right column). In order to construct the feature vector of a given sketch-gaze distance curve, we measure its similarity to each of these characteristics curves and use the degree of matching as an informative feature for identifying tasks. The first feature of our novel gaze-based feature representation corresponds to this vector of similarity scores.

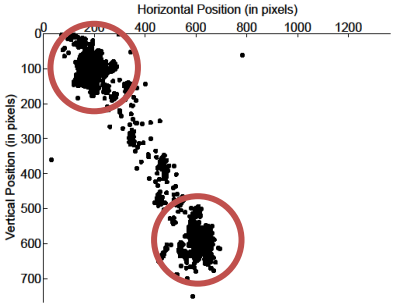
**Table 1. The column on the left shows pen-based virtual interaction tasks included in our research. The column on the right shows characteristic curves acquired from sketch-gaze distance curves of each task.**



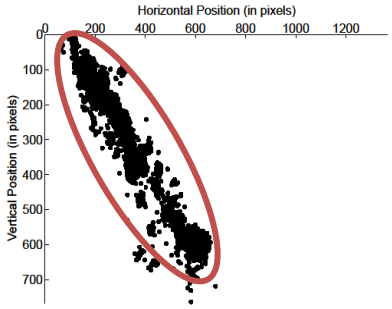

## 3.2 Feature 2: Within-Cluster Variance of Gaze Positions

As shown in Figure 4, eye gaze points collected during the execution of a task exhibit different clustering behaviors for different virtual interaction tasks. Based on this observation, we measure how the gaze points are clustered and spread out along the task path and use this measure as a useful and discriminative feature for identifying tasks.

In order to quantify the spatial distribution of gaze points along the path of a virtual interaction task, we measure the mean within-cluster variance of clustered gaze points for each task instance. We cluster the gaze points using MATLAB's *k-means clustering* algorithm and repeat this procedure three times for different $k$ values as $k = 1, 2, 3$. The second feature of our novel gaze-based feature representation corresponds to this vector of variance scores.




**a) Gaze data for the *drag* task – gaze points are clustered around the beginning and end points of the task trajectory resulting in gaze point clusters with low within-cluster variance.**




**b) Gaze data for the *free-form drawing* task – gaze points are scattered along the pursuit path resulting in gaze point clusters with high within-cluster variance.**

**Figure 4. Gaze data corresponding to 10 repeated task instances of a user.**

# 4. TASK PREDICTION AND EVALUATION

We assess the efficacy of the features presented above in predicting pen-based virtual interaction tasks based on the prediction accuracy. We compare the prediction accuracy of our novel gaze-based feature representation to that of various sketch-based feature representations that have been shown to work well for hand-drawn sketch data, namely IDM Features [3] and Zernike Moments [4]. In summary, our results suggest that the gaze-based feature representation is significantly better in capturing the richness and complexity of our user input when compared to commonly utilized and well-established sketch-based feature representations in the literature.

For all accuracy tests, we used the LIBSVM [5] implementation of Support Vector Machines. We measured the accuracies in line with the standard three-step machine learning pipeline, where first feature vectors are extracted from a set of data samples (Figure 5), then classifiers are trained using these feature vectors, and finally accuracies are measured using unseen data. Figure 6 summarizes the mean accuracies for individual feature representations.

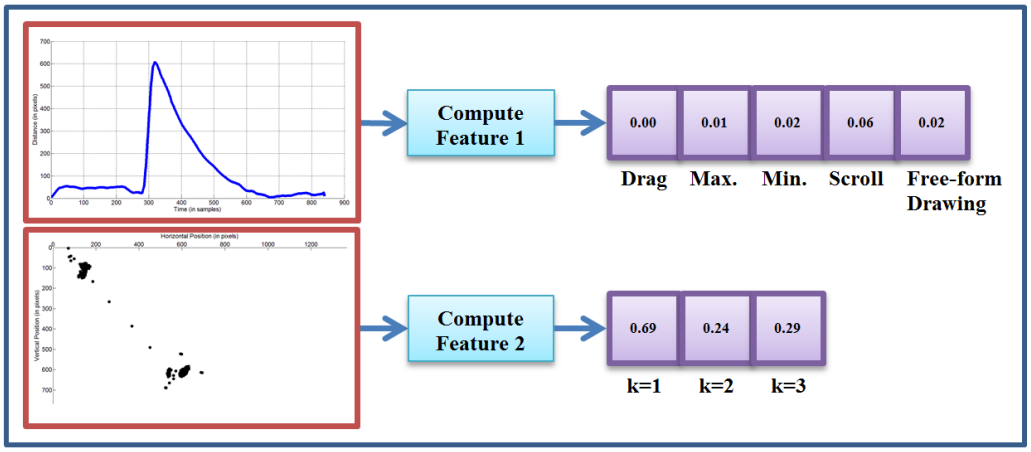



**Figure 5. Feature extraction in our framework. For each task instance, we compute a feature vector of size 8. This feature vector is then normalized by standardization.**

We conducted a two-way ANOVA to examine the effect of feature representation on prediction accuracy. ANOVA revealed a main effect of feature representation on prediction accuracy across the *Gaze-Based Features*, *IDM Features,* and *Zernike Moments* conditions at the $p < .05$ level, $[F(2,12) = 294.767, p < 0.001]$. Post-hoc comparisons using the Tukey HSD test indicated that the mean score for the *Gaze-Based Features* condition $(80.95 \pm 2.89)$ was significantly higher than the *IDM Features* condition $(58.57 \pm 3.20, p < 0.001)$ and the *Zernike Moments* condition $(37.55 \pm 2.31, p < 0.001)$.

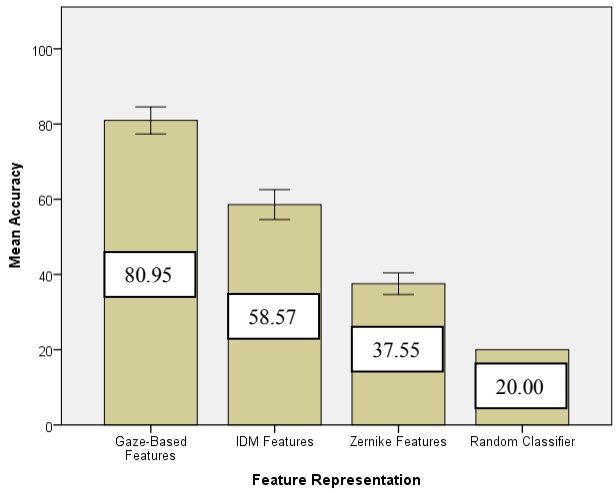



**Figure 6. Mean accuracy scores for each feature representation. Error bars indicate 95% confidence interval.**

## 5. RELATED WORK

We have presented a novel gaze-based system for predicting virtual manipulation tasks during pen-based interaction. Campbell et al. [6]'s work is one of the first examples of virtual task predictors in the literature. They classify reading, skimming, and scanning tasks by using a wide range of eye movement patterns.

This was largely followed by studies focusing on intention prediction, i.e. predicting whether the user intends to interact with the system or not during natural interaction. One example is by Bader et al. [7]. They use a probabilistic model to predict whether the user's intends to select a virtual object or not with 80.7% mean accuracy. A similar work is by Bednarik et al. [8]. They use SVMs to predict whether the user intends to issue a command or not with 76% mean accuracy. In both cases, the prediction tasks are examples of binary classification, where the baseline random classifier has an accuracy of 50%. Contrastingly, in our case, the baseline random classifier has an accuracy of merely 20%. Therefore, this fact should be taken into consideration when interpreting the stated accuracy scores and relevant classifier gains (i.e. the measurement of improvement over the random classifier).

To the best of our knowledge, there are only a small number of studies that take virtual task prediction one step further and aim at multi-class virtual task prediction instead of binary virtual task prediction. The work of Courtemanche et al. [9] is the first prominent example. The authors assert that their approach to activity recognition is the first one to incorporate eye movements. They discretize eye movements with respect to interface-specific AOIs and merge this information with information gained from keystrokes and mouse clicks input by the user during interaction. They use HMMs to predict which among the three Google Analytics tasks (i.e. evaluating trends in a certain week, evaluating new visits, or evaluating overall traffic) is currently being performed by the user with 51.3% mean accuracy. Recent work by Steichen et al. [10] comprises the second example. Their application area is graph-based information visualization and similarly they rely on interface- and graph-specific AOIs for feature extraction. They use Logistic Regression to predict which among the five information visualization tasks (i.e. retrieve value, filter, compute derived value, find extremum, or sort) is currently being performed by the user with 63.32% mean accuracy.

The superiority of our work over these two closely related studies is threefold. First, both of these studies are interface-dependent since their feature extraction mechanisms involve analyzing eye movements with respect to predefined AOIs. On the contrary, our features eradicate the need for possibly subject- and interface-specific preprocessing steps common in gaze-based systems. Second, both of these studies have highly specific and limited application areas, namely Google Analytics tasks and graph-based information visualization tasks. In contrast, our work can be applied in all areas that involve basic interaction tasks like dragging, resizing, and scrolling. Accordingly, the application areas of our task prediction system range from basic interfaces to more complex document or image editing software. Third, in terms of prediction accuracy, our task prediction system is

comparably more accurate, therefore making it a better candidate for practical use.

# 6. CONTRIBUTIONS AND FUTURE WORK

We have presented a virtual task prediction system that uses eye gaze movements to reduce dependency on explicit mode selection mechanisms in pen-based systems. Our task prediction system opens 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" [11]. It is widely accepted that intelligent mode selection mechanisms that provide low cost access to different interface operations will dominate new user interface paradigms [12].

We have three major contributions. First, we present a multi-modal dataset that consists of eye-gaze and pen input collected from participants completing various virtual interaction tasks (e.g., dragging, scrolling, minimizing, maximizing etc.) using a pen-based interface. 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. Our second contribution is a novel gaze-based feature representation that is capable of capturing the differences observed in eye gaze behavior during various virtual interaction tasks. Our third contribution is a novel gaze-based task prediction system that is based on this feature representation.

In the light of promising findings presented in this paper, an immediate follow-up to our work might involve conducting experiments to evaluate the robustness of our gaze-based feature representation to variations in scale. Another immediate extension might explore whether gaze-based and sketch-based feature representations can be combined by ways of feature-level or classifier-level fusion to increase prediction accuracy. Another possible direction might involve running feature selection tests to evaluate the relevance and redundancy of our gaze-based feature representation in comparison with the sketch-based feature representations in consideration. Lastly, additional feature selection tests can be run to assess the separate contributions of each of our two gaze-based features.

A more substantial extension might involve building an exhaustive taxonomy of pen-based virtual interaction tasks. Using WordNet, we have already rounded up a list of approximately 200 actions. We plan to categorize these actions with respect to user's major high-level interaction goal into four groups as translation, manipulation, selection, and search. More experiments will be needed to verify whether our task prediction system or a similar system inspired by our current findings generalizes well to our task taxonomy.

Another substantial extension might explore the feasibility of using our task 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 [13], 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.

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 task prediction system is fairly accurate (with a success rate of approximately 80%), inaccurate predictions are still 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 task 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.

# 7. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support and funding of TÜBİTAK (The Scientific and Technological Research Council of Turkey) under grant number 110E175 and TÜBA (Turkish Academy of Sciences).

# 8. REFERENCES

[1] Wacom Cintiq 24HD [Online image]. Retrieved July 20, 2014, from https://www.wacom.asia/cintiq24hd/gallery.

[2] Sakoe, H. and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 26*, 1 (1978), 43-49.

[3] Ouyang, T.Y. and Davis, R. A visual approach to sketched symbol recognition. In *Proc. IJCAI 2009*, Morgan Kaufmann Publishers Inc. (2009), 1463-1468.

[4] 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.

[5] Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 3 (2011), 1-27.

[6] Campbell, C.S. and Maglio, P.P. A robust algorithm for reading detection. In *Proc. PUI 2001*, ACM (2001), 1-7.

[7] Bader, T., Vogelgesang, M., and Klaus, E. Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In *Proc. ICMI-MLMI 2009*, ACM (2009), 199-206.

[8] Bednarik, R., Vrzakova, H., and Hradis, M. What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In *Proc. ETRA 2012*, ACM (2012), 83-90.

[9] Courtemanche, F., Aïmeur, E., Dufresne, A., Najjar, M., and Mpondo, F. Activity recognition using eye-gaze movements and traditional interactions. *Interacting with Computers 23*, 3 (2011), 202-213.

[10] Steichen, B., Carenini, G., and Conati, C. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proc. IUI 2013*, ACM (2013), 317-328.

[11] Nielsen, J. Noncommand user interfaces. *Communications of the ACM 36*, 4 (1993), 83-99.

[12] 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.

[13] Norman, D.A. *The design of everyday things.* Basic Books, 2002.