Dwell Selection with ML-based Intent Prediction Using Only Gaze Data

TOSHIYA ISOMOTO, University of Tsukuba, JAPAN
SHOTA YAMANAKA, Yahoo Japan Corporation, JAPAN
BUNTAROU SHIZUKI, University of Tsukuba, JAPAN

We developed a dwell selection system with ML-based prediction of a user's intent to select. Because a user perceives visual information through the eyes, precise prediction of a user's intent will be essential to the establishment of gaze-based interaction. Our system first detects a dwell to roughly screen the user's intent to select and then predicts the intent by using an ML-based prediction model. We created the intent prediction model from the results of an experiment with five different gaze-only tasks representing everyday situations. The intent prediction model resulted in an overall area under the curve (AUC) of the receiver operator characteristic curve of 0.903. Moreover, it could perform independently of the user (AUC=0.898) and the eye-tracker (AUC=0.880). In a performance evaluation experiment with real interactive situations, our dwell selection method had both higher qualitative and quantitative performance than previously proposed dwell selection methods.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; *HCI theory, concepts and models.*

Additional Key Words and Phrases: gaze-based interaction, user intent, hands-free, target selection, eye-tracker

ACM Reference Format:
Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. 2022. Dwell Selection with ML-based Intent Prediction Using Only Gaze Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 120 (September 2022), 21 pages. https://doi.org/10.1145/3550301

1 INTRODUCTION

1.1 Background

A widely used approach in gaze-based interaction is dwell selection [33–35], in which a user selects a target by looking at it for a particular duration (i.e., dwelling on it). The particular duration is called the "dwell-time," and dwell selection based on the dwell-time is referred to as DT selection. Generally, in DT selection, when the duration for which the gaze stays on a target exceeds the dwell-time, the system detects the dwell (we refer to this as DT detection) and determines the user's intent as "selecting the target." However, because visual information is perceived through the eyes, the user often spends time for other actions such as reading text. Accordingly, if DT selection occurs during such actions, it is an unwanted selection; this is called the *Midas-touch problem* [33–35].

To prevent the Midas-touch problem, many researchers have attempted to predict the user's intent. One approach is to adjust the dwell-time by making it larger or smaller according to the situation or the user (e.g., [45, 51, 60]). However, even a long dwell-time (e.g., 5 s) cannot prevent the Midas-touch problem when the user continuously looks at a target while thinking about something or observing the target. A more robust

Authors' addresses: Toshiya Isomoto, isomoto@iplab.cs.tsukuba.ac.jp, University of Tsukuba, Tsukuba, Ibaraki, JAPAN; Shota Yamanaka, syamanak@yahoo-corp.jp, Yahoo Japan Corporation, Chiyoda, Tokyo, JAPAN; Buntarou Shizuki, shizuki@cs.tsukuba.ac.jp, University of Tsukuba, Tsukuba, Ibaraki, JAPAN.

2474-9567/2022/9-ART120 $15.00
https://doi.org/10.1145/3550301

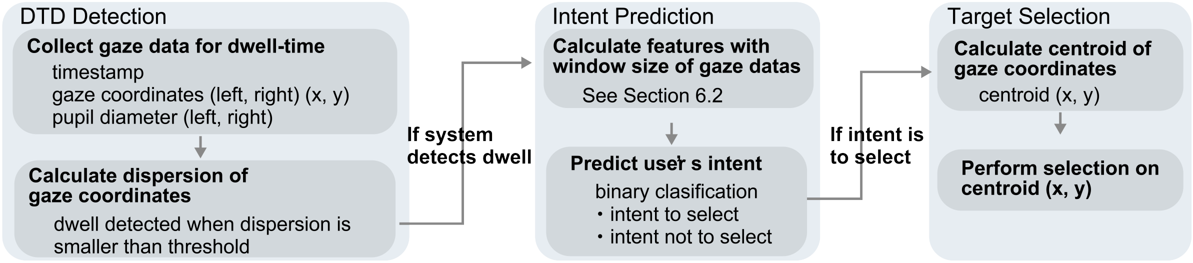

DTD Detection
Collect gaze data for dwell-time
timestamp
gaze coordinates (left, right) (x, y)
pupil diameter (left, right)
Calculate dispersion of gaze coordinates
dwell detected when dispersion is smaller than threshold
If system detects dwell
Intent Prediction
Calculate features with window size of gaze datas
See Section 6.2
Predict user' s intent
binary clasification
・intent to select
・intent not to select
If intent is to select
Target Selection
Calculate centroid of gaze coordinates
centroid (x, y)
Perform selection on centroid (x, y)

Fig. 1. Overview of DTD-ML selection.

approach is to use voluntary eye movements (e.g., saccades, pursuits, and vergences) to allow the user to look at a target for longer than the dwell-time. Using voluntary eye movements also helps predict a user's explicit intent for selection. However, researchers (e.g., [31, 41, 72]) have pointed out that voluntary eye movements can cause eye fatigue and that an additional UI is necessary for saccade-based and pursuit-based selection.

Instead, we explore the following research question: "Can intent prediction using eye-related information such as gaze movement, pupil diameter, and vergence prevent the Midas-touch problem when using a dwell as a user's action?" As there is a variety of eye-related information, information that represents the intent to select or not may be involved, which might allow us to predict a user's intent. However, identifying such information is not simple; thus, we adopt intent prediction based on machine learning (ML) to interpret the user's intent from the eye-related information. Our dwell selection thus consists of dwell-time and dispersion (DTD) based dwell detection and ML-based intent prediction, as shown in Fig. 1. We refer to our dwell selection method as DTD-ML selection.

1.2 Contributions of This Paper

- We show the DTD-ML selection that combines the advantages of both dwell selection (i.e., ease of selection) and voluntary eye movements (i.e., ease of intent prediction).
- We collect labels for creating an ML-based intent prediction model from five different tasks, which represent four interactive situations and one everyday situation without any manipulation, as described in Section 4.
- We show that our intent prediction achieves an area under the curve (AUC) of the receiver operator characteristic curve of 0.903; it also achieves high AUC values independent of the user and eye-tracking frequency, as described in Section 5.
- We show that DTD-ML can prevent 40.2% of unwanted selection in comparison to DTD selection, and that it has equal or better usability than both DT and DTD selection, as described in Section 6.

1.3 Motivation for Ubiquitous/Wearable Computing

In recent years, the development of devices equipped with eye-trackers has focused on wearable devices (e.g., smart glasses). In both cases, an interface presents information that is linked to the surrounding environment, with a view towards establishing ubiquitous computing. For example, when a user looks at a food, it is expected that the interface will present additional information, such as the food's calorie count and allergy status, according to the user's intent without additional explicit actions. Because DTD-ML is based on dwell selection, which would be the most implicit input among gaze inputs (e.g., gaze-gestures and gaze-combined multimodal interactions), an interface with DTD-ML presents information to satisfy the user's subconscious interests. Moreover, through the development of this interface for smart glasses in the future, our research will provide a bridge from gaze input (especially dwell selection) to ubiquitous computing and wearable computing.

2 RELATED WORK

We describe research on gaze-based selection and intent prediction by focusing on gaze-only interaction. While those are out of scope of this study, some researchers have proposed a combination of gaze-based interaction with other modalities such as touch [38, 55, 67], hand gesture [10, 56, 69], and voice [48]. This combination overcomes gaze-based interaction problems by using the second modality for accurate manipulation and detection of explicit intent. Moreover, gaze-based interaction can be enriched (thus overcoming its lack of functionality).

2.1 Selection with Dwell Detection

A user of DT selection can easily select a target by dwelling on it. The dwell is detected by measuring the duration for which the gaze coordinates are at the target and comparing that duration with the dwell-time. However, this simple detection process causes the Midas-touch problem.

The easiest solution is to use a longer dwell-time; however, this solution decreases usability. Researchers have thus sought to find solutions while keeping a shorter dwell-time. To achieve fast, robust DT selection, most researchers adjust the dwell-time depending on the situation. In dwell-typing research, the dwell-time is adjusted according to the probability of a key being typed [23, 45–47, 50, 57, 59, 60]. Another approach is to adjust the dwell-time according to the target [51] or the eye movement before landing on the target [30]. However, even though the task's cognitive load strongly affects the occurrence of the Midas-touch problem [77], those works used a selection task with colored targets or simple images.

Eye movements are typically not used for interacting with a computer but for perceiving information, paying attention, or expressing intent. Thus, researchers have used metrics other than the dwell-time.

2.2 Selection with Voluntary Eye Movements

Selection with a voluntary eye movement such as a saccade, pursuit, or vergence requires a two-step action: a fixation[1] to choose a target from among potential targets, and voluntary eye movements to select the target. A fixation is equivalent to landing a cursor on a target, and a voluntary eye movement is equivalent to left-clicking in mouse-based interaction. This two-step action enables detection of a user's explicit intent.

For selection with a saccade (i.e., a ballistic eye movement), an additional button is used to confirm the target [21, 44, 49, 52, 64, 66, 74]. For selection with a smooth pursuit (i.e., an eye movement that follows a moving object), an additional moving object is necessary [15, 17, 63, 71–73]. For selection with a vergence, the user first fixates on a target and then refocuses on a distant object for confirmation. A vergence needs no guidance; however, some researchers have used an additional UI (e.g., a physical object between the user and display) to facilitate selection [2, 40, 41].

Although voluntary eye movements effectively prevent the Midas-touch problem, the approach has disadvantages as compared with DT selection. One is fatigue due to the *voluntary* eye movements, because humans subconsciously perform saccades, pursuits, and vergences, yet they are rarely performed consciously. In contrast, DT selection is the least fatiguing among the gaze-based interactions, because we consciously move our eyes to "look" at something. Another disadvantage of voluntary eye movement methods is the need for an additional UI.

2.3 Intent Prediction for Gaze-based Interaction

Several researchers have used gaze data such as fixations, saccades, and pupil changes to detect attention (e.g., [4, 75]), cognitive states (e.g., [25]), and decision intent (e.g., [36, 42]). Such detection has been used not only for gaze-based interaction but also for designing a web page or visualizing a user's intent.

[1] The terms "fixation," "glance," and "dwell" have been used interchangeably, depending on the specific research topic; however, these detection methods are often not explained in detail. Thus, we treated them as the same term and called as "fixation."

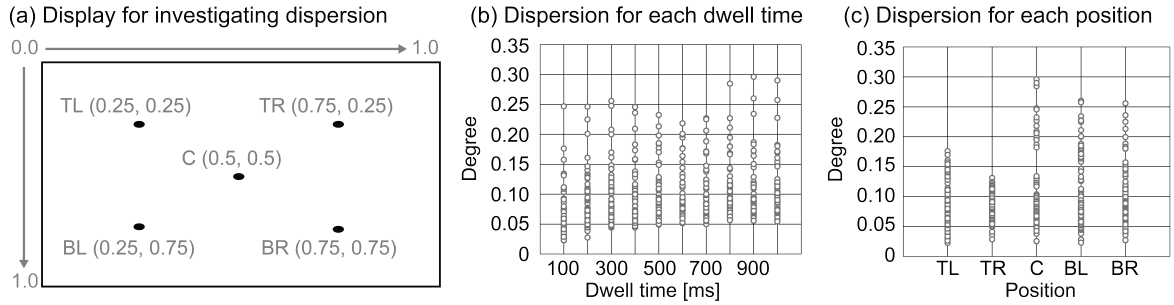


Fig. 2. Display used for investigating dispersion in the preliminary experiment to determine the dispersion threshold: (a) the points where the participants looked; and the dispersion results (b) for each dwell-time and (c) for each position.

Other researchers have explored applications of user intent prediction [9, 19, 22, 53] and developed ML-based intent prediction systems [5, 13]. In the ML-based systems, a positive class representing an intent to interact is mapped to a click with a controller, while a negative class is mapped to a fixation that is not associated with a click. For example, gaze data without clicking for a duration of 200 ms was used as a negative class [13]. The ML-based system in [5] was evaluated for an eight-tile puzzle game. However, it used gaze data during one fixation after a button press event (i.e., selection of a tile) as a positive class; thus, the system performed inadequately in real time (as pointed out in [13]). In contrast to the previous research, to achieve higher intent prediction performance, we adopted multiple experimental tasks and labeled user's intent with positive and negative classes.

In another intent prediction approach, Salvucci and Anderson predicted a user's desired target via a probabilistic algorithm using gaze coordinates and task context to overcome another problem of gaze-based interfaces, namely the lack of eye-tracking accuracy [61]. Since the eye-tracking generally involves noise, such target prediction, in combination with our user intent prediction, is also helpful in establishing gaze-based interaction.

3 OUR DWELL SELECTION (DTD-ML SELECTION) SYSTEM

Fig. 1 illustrates how our system detects a dwell and predicts the user's intent.

3.1 Dwell-time and Dispersion-based (DTD) Dwell Detection

In our system, DTD detection contributes to a rough screening of the user's intent to select and triggering ML-based intent prediction. The DTD detection system detects a dwell if the dispersion during the dwell-time is less than a dispersion threshold. Because of the dispersion threshold, the user needs to dwell more intentionally than in DT selection, but this helps prevent the Midas-touch problem. This concept is the same as the I-DT fixation detection algorithm [62].

In this study, we determined the dwell-time and dispersion threshold from a preliminary investigation since there has been no detailed investigation of suitable thresholds, although DTD detection has been used in commercial software [29, 65, 66, 68] and for other interactions [16, 27, 28, 31, 37, 70]. Specifically, we investigated how a user's gaze varies while intentionally dwelling on a point. Fourteen male volunteers (aged 21–25) participated. We used a Tobii Eye Tracker 4C (sampling rate: 90 Hz) with a pro license for research; we attached this to the bottom of the 24-inch (1980×1080 pixels) non-glare display. The participant's head was positioned approximately 65 cm from the display. We asked the participants to calibrate the eye-tracker before starting the first task. For each task, they looked at each of five points on a display, as shown in Fig. 2a, for 2,000 ms. We collected 70 attempts (14 participants × 5 points) in total.

We first eliminated eight attempts that included a saccade with the I-VT algorithm whose velocity threshold was 100°/sec [62]. To have stable gaze data, we used the last 1,000 ms of gaze coordinates from the remaining 62 attempts to calculate the thresholds. We then calculated the standard deviation of 10 dwell-times (100, 200, 300, ..., 900, 1,000 ms. If the dwell time is 100 ms, we used first 100 ms (i.e., 1,000 ms to 1,100 ms out of 2000 ms.)) at the gaze coordinates as the dispersion. The results showed that all dispersion were less than 0.3°, independent of the dwell-time and the position (Fig. 2). We thus used 0.3° as the dispersion threshold and 600 ms as the dwell-time in our system. We chose 600 ms for two reasons: first, it is not a long dwell-time as compared with those in previous studies; second, it is an appropriate dwell-time for the cognition model [32]. Tuning these thresholds for the user, position, and other aspects such as the task and familiarity with gaze input would further clarify the user's intent, and this is left for future work.

3.2 Intent Prediction with an ML Model

After dwell detection, the system predicts the intent to select or not by using an ML model. The system first calculates features from the window size (2,000 ms in this paper, as described in Section 5.4.5) of gaze data before a dwell is detected. For the features, we use eye-related information (saccades, fixations, vergences, and pupil changes) and quantitative data (eye movement distances and durations), that we describe in detail in Section 5.2.

Note that our system only predicts two intents: intent to select and intent not to select. If we predicted various other intents (e.g., attention, gaining more information), we could develop further interactions. Here, however, we focus on binary prediction to establish a gaze-based intent prediction system for future development.

3.3 Target Selection

If the predicted intent is to select, we calculate the centroid of the gaze coordinates, $\mathrm{C}_{x/y}$, during the dwell. The system then activates selection to $\mathrm{C}_{x/y}$.

4 EXPERIMENT 1: LABELING OF USER'S INTENT

In this experiment, we collected ground-truth labels representing the user's intent to select or not to select.

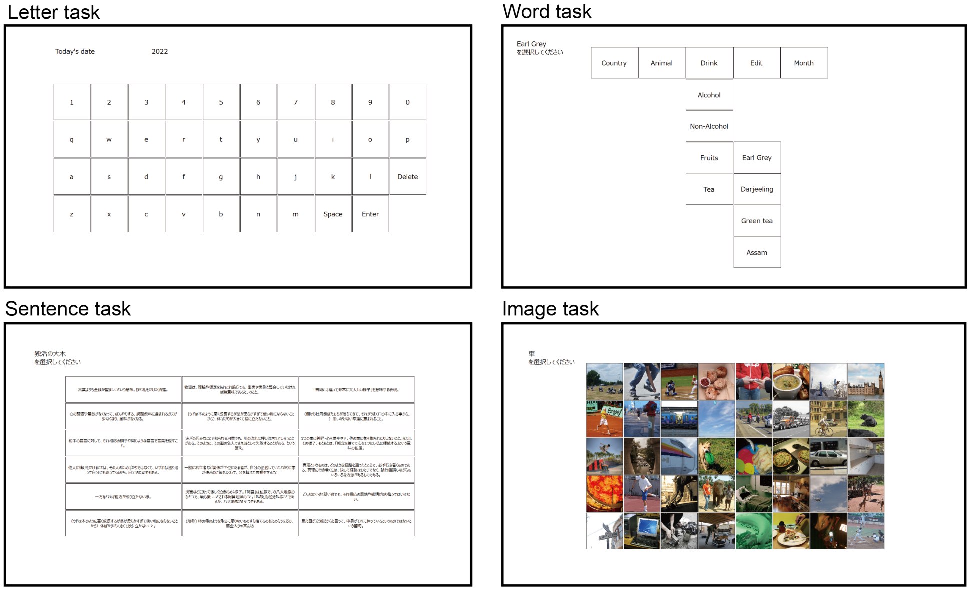
4.1 Participants and Apparatus

We recruited 24 university students (five females and nineteen males, all Japanese) aged 20–26 ($M = 22.9$). Fifteen of them had participated in an experiment using an eye-tracker. Each received JPY 5,000 (~USD 45).

We used the Tobii Pro Spectrum and Tobii Pro Fusion as eye-trackers; both were attached to the bottom of the 24-inch (1980×1080 pixels) non-glare display. One reason for using two different eye-trackers was that it was necessary to investigate whether we could use our ML model with different eye-trackers, as the eye-tracking frequency generally differs from one device to another. Another reason for using two different eye-trackers was that some participants could not calibrate the Tobii Pro Fusion, apparently because of the incompatibility of pupil detection for Asians[2].

Twelve participants used the Tobii Pro Spectrum at 1,200 Hz, eight used the Tobii Pro Fusion at 250 Hz, and four used the Tobii Pro Spectrum at 120 Hz; most commercial eye-trackers sample gaze data at 120 Hz (e.g., the Tobii Eye Tracker 5 and HTC VIVE PRO EYE). The participant's head was positioned approximately 65 cm from the display. The participant used a keyboard to control the task. The experiment was conducted in a room with fluorescent light at approximately 810 lux.

[2] For the details of the pupil detection method, see https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/what-is-dark-and-bright-pupil-tracking/. From communication with staff at Tobii, we decided to use the Tobii Pro Spectrum.



独活の大木
を選択してください

言葉よりも金銭が望ましいという意味。証と札をかけた洒落。	物事は、理論や仮定をあれこれ論じても、事実や実例と整合していなければ無意味であるということ。	「普段とは違って非常に大人しい様子」を意味する表現。
心の緊張や意欲がなくなって、ぼんやりする。炭酸飲料に含まれるガスが少なくなり、風味がなくなる。	（ウドは木のように高く成長するが茎が柔らかすぎて使い物にならないことから）体ばかりが大きくて役に立たないこと。	（棚から牡丹餅ぼたもちが落ちてきて、それがうまく口の中に入る事から。）思いがけない幸運に恵まれること。
相手の暴言に対して、それ相応の調子や同じような暴言で言葉を返すこと。	泳ぎの巧みなことで知られる河童でも、川の流れに押し流されてしまうことがある。そのように、その道の名人でさえ時として失敗することがある、という譬え。	1つの事に神経・心を集中させ、他の事に気を取られたりしないこと。またはその様子。もともとは、「雑念を捨てて心を1つにし仏に帰依する」という意味の仏語。
他人に情けをかけることは、その人のためばかりではなくて、いずれは巡り巡って自分にも返ってくるから、自分のためでもある。	一般に若年者など関係が下位にある者が、自分の企図していたとおりに事が運ぶのに気をよくして、分を超えた言動をすること	真理というものは、どのような経路を通ったところで、必ず行き着くものである。真理に行き着くには、決して経路はひとつでなく、試行錯誤しながらもいろいろな方法があるものである。
一方をとれば他方が成り立たない様。	災害などにあって激しく泣きわめく様子。「阿鼻」は仏教でいう八大地獄のひとつで、最も厳しいとされる阿鼻地獄のこと。「叫喚」は泣き叫ぶことであるが、八大地獄のひとつでもある。	どんなに小さく弱い者でも、それ相応の意地や感情があり侮ってはいけない。
（ウドは木のように高く成長するが茎が柔らかすぎて使い物にならないことから）体ばかりが大きくて役に立たないこと。	（俗称）柿の種のような取るに足りないものすら捨てるのをためらうほどの、筋金入りの吝ん坊	見た目が立派だからと言って、中身がそれに伴っているというものではないという警句。

Fig. 3. Displays used for the tasks.

4.2 Tasks

Because the eye movement and pupil diameter vary with the action, environment, and visual stimulus, we collected labels and gaze data from five different tasks: a letter task, a word task, a sentence task, an image task, and a movie task. These tasks represent four interactive situations (selecting a letter, word, sentence, or image) and one everyday situation without any intent to select (watching a movie).

The participants selected the target appropriate to each instruction by using DTD selection with a 600 ms dwell-time and a 0.3° dispersion threshold. We asked them to intentionally dwell on a point in the object rather than looking at it peripherally. For example, to select the sentence "This is a pen," we asked them to pick one letter (e.g., "p") and dwell on it. Similarly, to select an image of a dog, they picked a point in the image (e.g., the nose) and dwelled on it. Because we asked them to label the positive class when they performed target selection, we adopted this instruction to unify the action of "intentional dwell" in this experiment.

Fig. 3 shows the display used for each task. We determined the size of the target at which the eye-tracking performance (i.e., the offset and precision) did not affect selection. The participants read the task instruction and then pushed the space key to move on. Regardless of the participant's intent, the system displayed the labeling form, which contained a questionnaire regarding the intent and no content when it detected a dwell. For all tasks, to eliminate any potential side effects, we did not give the participants visual feedback; however, they could recognize that a dwell was detected through the labeling form's appearance, except in the movie task.

4.2.1 *Letter Task.*

The participants successively selected keys on a displayed keyboard. The size of each key was 3.5°×3.5°. The keyboard consisted of 10 digits, qwerty-arranged keys, a space key, a delete key, and an enter key. The task consisted of typing the date and the participant's name, age, and hobby; e.g., one instruction was "write today's date." There was no specific format, so the participants could enter the date freely, e.g., "20210801" or "0801." They finished a trial by selecting the enter key, and then they labeled their intent for each key selection.

We assume that this task represents a situation in which the user selects a letter; selection of one of four choices is another possible situation.

4.2.2 Word Task. The participants manipulated a three-layer hierarchical menu and selected an item written in word(s). The size of each item was 4.5°×3.0°. The participants performed 20 selections for randomly chosen instructions. After selecting an item in the third layer, they moved on to the next instruction. For example, for the instruction "select Japan," the participants selected "Country" → "Asia" → "Japan." We asked the participants to search for an appropriate target as much as possible; if they could not find one, we asked them to select an arbitrary target. We did not limit the number of times opening the menu or the time to select the target. The participants labeled their intent for each menu item selection.

We assume that this task represents a situation where the user selects a word; directory manipulation and item selection are other possible situations.

4.2.3 Sentence Task. We asked the participants to select appropriate Japanese meanings (sentences) for idiomatic phrases (instructions). We used 100 pairs of phrases and meanings[3]. The size of each sentence was 11.0°×2.5°. Each participant performed 30 selections for randomly chosen phrases. From the 100 pairs, we arranged 18 choices, consisting of one correct meaning and 17 randomly chosen meanings, in a 3×6 grid. We asked the participants to select the correct choice as much as possible; if they could not find or did not know the meaning, we asked them to select the most plausible choice. They labeled their intent for each selection.

We assume that this task represents a situation in which the user reads sentences and selects one, such as a hyperlink on a web page.

4.2.4 Image Task. We asked the participants to select an image that was appropriate for a given verbal instruction. We used a set of 64,332 images[4]. The size of each image was 3.5°×3.5°. Each participant performed 100 selections for randomly chosen instructions. We arranged 40 choices consisting of at least one correct image along with randomly chosen images in an 8×5 grid. We asked the participants to select the correct image, but if they could not find it, we asked them to select the most plausible image. They labeled their intent for each selection.

We assume that this task represents a situation in which the user searches for an image and selects it, such as an image search on a web page or icon selection in a desktop window.

4.2.5 Movie Task. As opposed to the other tasks, we told the participants that it was unnecessary to select a target and instead asked them to watch a movie as if they were watching it on YouTube or Netflix. We used 500 movies from ActivityNet [18]; we streamed them by using a full-screen mode of Windows Media Player without a UI. Each participant watched movies for 10 minutes. They were allowed to be lost in thought if a movie was not attractive.

Even though this task simply plays a movie on a desktop computer, the gaze data collected through it involved various kinds of information. For example, because we chose the movies regardless of the participants' interests, we could collect various kinds of intent, attention, and interest depending on the movie, content and period. Similarly, the direction, distance, and duration of users' gaze movements, saccades, and fixations varied. We thus conducted this movie task to collect negative data in the form of dummy data representing gaze data that did not involve intentional manipulation in daily life.

4.2.6 Intent labeling. The participants gave their intent with respect to dwell detection with a physical keyboard in accordance with the guidelines listed in Table 1. In the following analysis, we used the detected dwells that were labeled "Yes" as the positive class and those that were labeled "No" as the negative class. A selection labeled as the negative class was treated as an unwanted selection. Note that there was no selectable UI for the movie

[3] From https://www.wiktionary.org/, licensed under CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/)

[4] From https://visualgenome.org/, licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Table 1. Guidelines for intent labeling

Situation in which labeling form was displayed	Labeling guideline
Intentionally dwelling on point in correct target	Yes (positive class)
Intentionally dwelling on point not in correct target	Yes (positive class)
Correct target viewed before form was displayed, but participant still thinking about target's correctness	No (negative class)
Participant thinking, searching, or lost in thought	No (negative class)

Table 2. Numbers of labeled classes.

Task	Letter	Word	Sentence	Image	Total of left 4 tasks	Movie
Positive class ([%])	788 (99.12)	1,474 (93.23)	425 (59.03)	2,053 (85.54)	4,740 (86.34)	None
Negative class ([%])	7 (0.88)	107 (6.77)	295 (40.97)	347 (14.46)	756 (13.76)	10,586 (100.0)

task, and the participants did not label their intent; accordingly, we labeled all the detected dwells in the movie task as negative classes.

4.3 Procedure

We asked each participant to calibrate the eye-tracker before starting the first task. The task order was randomized among the participants. They were allowed to take an optional break when the instruction form was displayed. The experiment took an average of 68 minutes per participant.

4.4 Labeling Results

We list the labeling results in Table 2. Even without the ML-based intent prediction, there were fewer negative classes for the letter task than for the other tasks. This result would be due to the letter task having a lower cognitive load than the others. The negative class percentage was the highest for the sentence task. Many participants said that the sentence task was challenging because it was difficult to find the correct meaning of the Japanese phrase; thus, much time was spent reading and thinking about the sentence. Accordingly, the system might have mistakenly detected dwells, which the participants labeled as negative classes. For the movie task, there was a total of 10,586 negative classes, which suggests that there are many possibilities for incorrect detection of dwell during everyday situations.

Note that, balancing the number of labels for each task is preferable; however, in this study, we used imbalanced labels to create an ML model that was robust against both false negatives and false positives. For example, if we ignored data collected from the letter task, whose labels were biased positively, the prediction may have caused false negatives. In contrast, if we ignored data collected from the movie task, whose labels were biased negatively, the prediction may have caused false positives. Given the trade off between the Midas-touch problem and the ease of selection, we decided to use both positively and negatively biased data.

5 ML MODEL FOR INTENT PREDICTION ON DWELL SELECTION

We used the results of EXPERIMENT 1 to create an ML model for intent prediction.

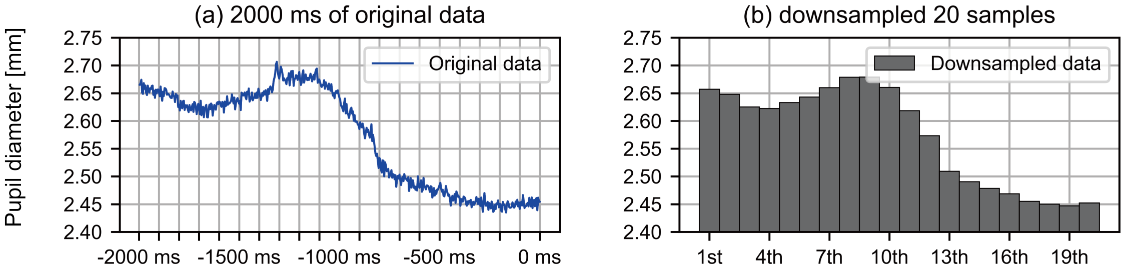


Fig. 4. Example of raw data for (a) **pupil** and (b) its downsampled values.

5.1 Data Processing

As listed in Table 2, the positive-to-negative class ratio was unbalanced for each task. Accordingly, we used the negative classes for the movie task to alleviate the imbalance. Specifically, to make the ratio 50:50, we randomly chose negative classes from the movie task for each participant.

To calculate features, we used 2,000 ms as the window size of gaze data before a dwell was detected; the details are described below in Section 5.4.5. The gaze data were the x/y coordinates ([0.0 (top left) −1.0 (bottom right)]) on the display, the pupil diameter ([mm]), and the timestamp. These data were collected for both the left and right eyes. For each timestamp, we calculated the average of the left and right pupil diameters (**pupil**), the averages of the x/y coordinates for the left and right eyes (**x** and **y**), and the difference between the x coordinates of the left and right eyes ($\mathbf{diff_x}$). We then downsampled these values to 20 values, i.e., the average values for every 100 ms of gaze data. Fig. 4 shows an example for **pupil**. Next, we calculated the changes between the last (20_{th}) value and each i-th ($i = 1, 3, ..., 19$) value (19 changes). Using the changes (instead of the original values) allowed us to eliminate the gaze data dependence on the user, environment, and task. We adopted this process to observe how the gaze data changed over 2,000 ms rather than in a short span (e.g., every 0.833 ms for 1,200 Hz), because gaze data do not change within a short span [12], and eye-tracking data contains noise. Moreover, we adopted the downsampling to cover the difference in the eye-tracking frequencies; this process helped us create a general ML model that was independent of the eye-tracking frequencies.

In addition, we used the I-VT algorithm [62] to detect fixations and saccades from the original x/y coordinates. We used 10°/sec for fixation detection and 100°/sec for saccade detection. Moreover, to exclude eye-tracking noise, we used 100 ms as the minimum duration of a fixation and 30 ms as the minimum duration of a saccade.

5.2 Features

We used the following gaze data to calculate the features listed in Table 3.

x and **y:** Changes in the x/y coordinate values indicate how the gaze moved during the 2,000 ms before dwell detection. Using changes gave more independent information than using an absolute gaze position on the display.

$\mathbf{diff_x}$**:** Changes in $\mathbf{diff_x}$ indicate whether the focus moved from or to the display (i.e., whether a vergence occurred) during the 2,000 ms before dwell detection. Although we could have determined how far the focus was from the display if we used the original values of $\mathbf{diff_x}$, the eye-tracking accuracy and the individual's eyesight may have affected the values. Thus, we used the changes in $\mathbf{diff_x}$.

pupil: Changes in **pupil** indicate how the user's interest, emotions, or awareness shifted during the 2,000 ms before dwell detection. We used the changes in **pupil** because the original values depended on the individual and the brightness of the location and the display.

Table 3. Calculated features. In total, we used 127 (= 80 + 35 + 12) features for ML.

Features		Numbers
plus, minus, absolute, and all (19) values of changes in **x**, **y**, **diff**$_{\mathbf{x}}$, and **pupil**	average, standard deviation (SD), amplitude, skewness, kurtosis	80 (4×4×5)
durations of saccades, durations of fixations, distances of saccades, distances of fixations, velocities of saccades	average, first value, last value, last value minus first value, minimum value, max value, amplitude	35 (5×7)
Changes in **x**, **y**, **diff**$_{\mathbf{x}}$, and **pupil**	1st value, 19th value, difference between 19th and 1st values	12 (4×3)

Table 4. AUCs of our intent prediction. The values except for *all* are averages. The important results in terms of this work's contributions and limitation aspects are highlighted in red and blue, respectively.

all	*all* (hyper-parameters)	*each-XXX*			*leave-one-XXX-out*		
		participant	task	frequency	participant	task	frequency
0.903	0.905	0.893	0.964	0.909	0.898	0.601	0.880

saccades and **fixations:** In addition to **x** and **y**, saccades and fixations indicate how the user's attention shifted during the 2,000 ms before dwell detection.

The features in the first and second rows of Table 3 are consistent with those used in previous works [5, 13]. Because these statistical values summarize the original data and would allow the prediction model to focus on its important characteristics, the prediction result would likely be better than using original data. In general, the changes' directions are important: for example, when we read a sentence, the gaze moves from left to right, resulting in positive changes in this environment. Thus, we calculate these statistical values with signs. In addition, we use the features in the third row because the first and last (19th) values and their differences represent how the data changes. These features are promising for determining the user's intent; still, it is complicated to decide the thresholds for each feature, and we thus use ML-based prediction.

5.3 Metrics for Evaluation

We used the area under the curve (AUC) of the receiver operator characteristic curve [7] as the primary metric for evaluating the prediction performance. A higher AUC value means a greater chance of achieving both a high true positive rate (TPR) and a high true negative rate (TNR), thus helping our prediction system deal with the trade off between the Midas-touch problem and the ease of selection.

5.4 Creating ML Model

We created prediction models for all data (*all*), the participants (*each-participant* and *leave-one-participant-out*), the tasks (*each-task* and *leave-one-task-out*), and the eye-tracking frequencies (*each-frequency* and *leave-one-frequency-out*), and we tested each of the models.

For *each-XXX*, we split the classes for one participant, task, or frequency into training, validation, and test data. For *leave-one-XXX-out*, we used the classes for one participant, task, or frequency as the test data, and we split the remaining classes into training and validation data. We performed five-fold cross-validation for training, validating, and testing the models. For the classifier, we used LightGBM, because it gave AUC values that were higher than those of the other classifiers that we tested (see Section 5.4.6).

5.4.1 *Overall Prediction Results.*

Table 4 summarizes the prediction results. For *all*, the AUC, accuracy, recall, precision, F1, and Matthews correlation coefficient (MCC) were 0.903, 0.826, 0.839, 0.818, 0.828, and 0.652, respectively. We calculated the TPR and TNR values with respect to the prediction probability threshold. They intersected at a value of 0.825, where the threshold was 0.524. With 0.80 as the threshold, we could achieve a TNR of 0.900, while the TPR fell to 0.696. Accordingly, as with the dwell-time, there is a trade off between the TPR and TNR.

We also give the prediction results obtained with hyper-parameters that we determined by using LightGBM Tuner from Optuna [3]. The tuned parameters were "lambda_1": 6.25e-06; "lambda_l2": 4.07e-06; "num_leaves": 28; "feature_fraction": 0.4; "bagging_fraction": 0.75; "bagging_freq": 5; and "min_child_samples": 20. For *all* with these hyper-parameters, the AUC, accuracy, recall, precision, F1, and MCC were 0.905, 0.829, 0.845, 0.819, 0.832, and 0.659, respectively.

5.4.2 *Prediction Results for Participants.*

The AUC values were high for both *each-participant* and *leave-one-participant-out*: they averaged 0.894 [0.802–0.967] and 0.898 [0.839–0.963], respectively. These results demonstrate that the model can predict the user's intent and is thus useful as a general model independent of the user. Given the limited diversity of the participants, their small age range might have resulted in the high AUC values. On the other hand, because we did not use the original values for **pupil** and $\mathbf{diff_x}$, which vary by individual, as features, similar results should be achievable for users with different attributes.

5.4.3 *Prediction Results for Tasks.*

A high average AUC value of 0.964 [0.898−0.994] was achieved for *each-task*; however, the value was 0.601 [0.443−0.703] for *leave-one-task-out*. Although we used the changes in the gaze data, they still depended on the task, and thus the AUC values for *leave-one-task-out* were not sufficient to make predictions, especially for the sentence task, whose AUC value was 0.443.

Because the movie task had one class, we did not create a prediction model for *each-task* for the movie task. As for *leave-one-task-out*, we trained the model with the classes of the letter, word, sentence, and image tasks. Before training, we downsampled the positive classes of those four tasks to equalize the class ratio. Testing gave a TNR of 0.463 when the prediction probability threshold was 0.5. With a higher threshold of 0.9, the TNR was 0.914. The high AUC for *each-task* and low AUC for *leave-one-task-out* show the importance of using more various tasks when creating a gaze-based intent prediction model.

5.4.4 *Prediction Results for Frequencies.*

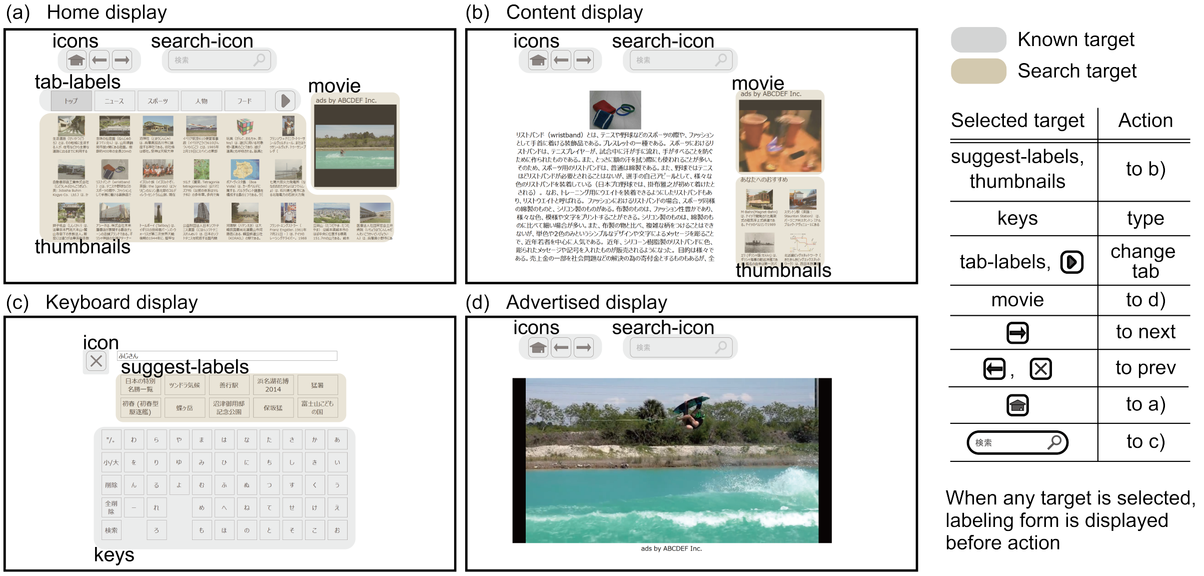
The AUCs for both *each-frequency* and *leave-one-frequency-out* were high, with respective averages of 0.909 [0.895−0.917] and 0.880 [0.859−0.902]. These results indicate the validity of the features used for the model with eye-trackers having different frequencies. However, eye-trackers mounted on an HMD and different eye-tracking (or pupil-tracking) methods may produce different results.

5.4.5 *Prediction Results for Window Size.*

We examined the prediction results for *all* with features that were created using window sizes for the gaze data of 600−2,900 ms, in 100 ms steps. The metrics increased with the window size: for example, the AUC value was 0.773 at 600 ms, 0.845 at 1,000 ms, 0.877 at 1,500 ms, 0.903 at 2,000 ms, 0.923 at 2,500 ms, and 0.934 at 2,900 ms. Although larger window sizes should be investigated, we could not do so because some of the gaze data collected within a task was shorter than 3,000 ms. Thus, when we used 3,000 ms as the window size, approximately 20% of the tasks were eliminated as compared to using 600 ms as the window size. Another issue is that a larger window size may cause overfitting for these tasks, with regard to display designs or target alignments. According to these results, we created features by using a window size of 2,000 ms, which was the smallest one that achieved an AUC value greater than 0.900.

5.4.6 *Prediction Results for Other Classifiers.*

We examined the prediction results for *all* with various classifiers: support vector machine, random forest, logistic regression, and LightGBM. The AUC values were 0.781, 0.825, 0.781, and 0.903, respectively. We thus used LightGBM as the classification algorithm, as mentioned above.



Selected target	Action
suggest-labels, thumbnails	to b)
keys	type
tab-labels, ▶	change tab
movie	to d)
➡	to next
⬅, ☒	to prev
🏠	to a)
検索 🔍	to c)

When any target is selected, labeling form is displayed before action

Fig. 5. Interface used in Experiment 2.

5.4.7 Use of Task- and Participant-dependent Gaze Data. We did not use the original values in the gaze data because they depended on the user, environment, and task. For example, if the interface design differs from that in Experiment 1, these values, especially **x** and **y**, will differ. Moreover, the original values of **pupil** depend on the light conditions or the type of visual stimulus [26]. While the use of those values increased the AUC values for *all* (>0.940), they could have caused overfitting that could not be displayed in the prediction test with the current data.

5.4.8 Feature Importance. The top 10 gains among the features were the average and kurtosis of the absolute values of **x** and **y**, the amplitude of all values of **x**, **y**, and **pupil**, the standard deviation and average of the plus values of **pupil**, and the last value of **pupil** (these scores are shown in the Supplementary). This result suggests the importance of how the gaze moves and how the pupil changes. Notably, the plus values of **pupil** and the last value of **pupil** seemed to have a large impact, because the diameter increases with the interest or emotion [26].

6 EXPERIMENT 2: PERFORMANCE EVALUATION

We tested how DT, DTD, and DTD-ML selection worked in a real interactive situation. In particular, we focused on how the dispersion threshold screened the user's intent and how the ML model predicted the intent.

6.1 Participants and Apparatus

We recruited 12 university students (four females and eight males, all Japanese) aged 20–24 (M =22.9). Six of them had participated in Experiment 1, and nine had participated in an experiment with a gaze-based interface. This experiment used the same apparatus and environment as in Experiment 1. Specifically, we used the Tobii Pro Spectrum at 1,200 Hz as the eye-tracker in Experiment 2.

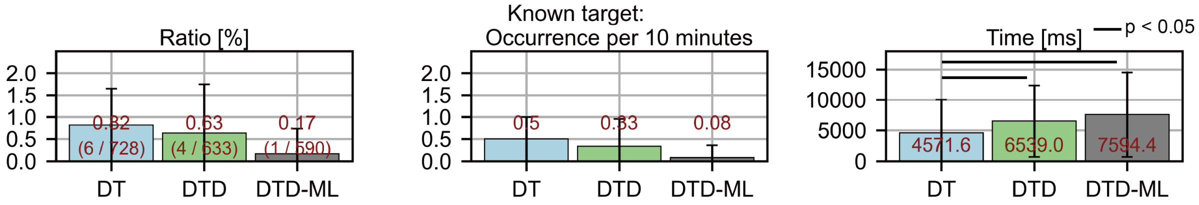


Fig. 6. Quantitative results for selection of known targets. The values in parentheses indicate the numbers of unwanted selections and total selections.

6.2 Task

The task was to interact with a dictionary-like interface, shown in Fig. 5, by using dwell selection. We roughly classified targets in the interface into two types. One was a *known target*, of which the participants knew the location and content; the other was a *search target*, which the participants had to search for or understand the content. We used keys, tab-labels, icons, and a search-icon as known targets, because their locations and contents remained the same throughout the experiment; other targets (i.e., thumbnails, movies, and suggest-labels) were used as search targets. The sentences and images in the target contents were taken from Wikipedia[5], while the movies were the same as in Experiment 1. When any target was selected, the labeling form was displayed. The participant gave their intent for selection as in Experiment 1.

The target sizes are 2.0°×2.0° for keys and icons, 4.0°×2.0° for tab-labels and suggest-labels, 4.0°×4.0° for thumbnails, 8.0°×4.0° for a movie, and 10.0°×2.0° for a search-icon. We determined these sizes by choosing a minimum target size and enlarging other targets appropriately for their meaning. We chose the minimum size as 2.0°, which was approximately 2.3 cm on the screen used here, making the size similar to that suggested in [20] (for filtered data, a target size of 1.9 cm×2.35 cm enables reliable interaction for at least 75% of users).

We used a dwell-time of 600 ms and a dispersion threshold of 0.3°. The window size was 2,000 ms. The prediction threshold was 0.800. We used the same ML model that gave the results for *all* (hyper-parameter) shown in Section 5.4.1.

6.3 Procedure

We asked each participant to calibrate the eye-tracker before starting the task. The order of the selection methods was randomized. We asked the participants to searching for a target whose content was attractive and to select that target. We did not limit the way of search and told them to freely interact with the interface. For each selection method, we asked the participants to interact for 10 minutes. We did not calibrate or adjust the ML model for each participant, nor did we allow the participant to train each selection method.

After the 10-minute interaction, the participants answered the System Usability Scale (SUS) [8] and the NASA-TLX [24] tests. Then, they rested for at least five minutes before moving to the next method. The experiment took an average of 53 minutes per participant. Each received JPY 5,000 (~USD 45).

6.4 Results

6.4.1 Quantitative Results.

For quantitative measures, we used the ratio of unwanted selections, the occurrence of unwanted selections, and the time to search for a target. The ratio was calculated from the number of selections labeled as "No" and the number of total selections. The occurrence was calculated from the number of total selections. The time was calculated by subtracting the time at which the labeling form closed from the time at

[5] https://en.wikipedia.org/, licensed under CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/)

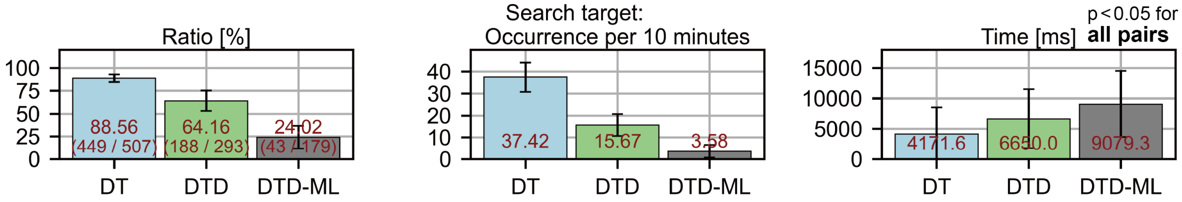


Fig. 7. Quantitative results for selection of search targets. The values in parentheses indicate the numbers of unwanted selections and total selections.

which a target was selected. Fig. 6 and Fig. 7 show the results for selecting the known targets and search targets, respectively. For DT, DTD, and DTD-ML selection, the ratio and occurrence decreased in that order, whereas the time increased in that order, regardless of the target.

To compare the three selection methods, we used the Friedman test ($\alpha = 0.05$) and the Bonferroni correction test ($\alpha = 0.05$) for the ratio, occurrence, and time. We found significant differences for the ratio and occurrence for search targets. This indicates that the screening of user intent with DTD detection and the intent prediction with an ML model works well; DTD-ML selection (ratio: 24.02) prevented 40.2% of unwanted selection in comparison to DTD selection (ratio: 64.16), and DTD selection prevented 24.4% in comparison to DT selection (ratio: 88.56). For known targets, there were no significant differences for the ratio and occurrence. This result confirms both the usefulness of DT selection for known targets and the result of the letter task in Experiment 1. As for the time, there were significant differences between DT and the other selection methods for both known and search targets. Both DTD and DTD-ML selection allowed the participants to search for a target more carefully. However, this result also suggests that DT selection allows faster selection than DTD and DTD-ML selection.

For the ratio, occurrence, and time with DTD-ML selection, there was no significant difference between the participants who did and did not participate in Experiment 1. Because we used the ML-based intent prediction model created via Experiment 1, this result validates the model's user-independence.

6.4.2 Qualitative Results. We show the NASA-TLX and SUS results in Fig. 8 and Fig. 9, respectively, for each selection method. We tested significant differences in the scores of the three selection methods with the same Friedman and Bonferroni correction tests.

The averages and ranges of the overall NASA-TLX scores were 47.5 [32.33–58.0], 24.58 [14.67–33.0], and 20.31 [11.33–29.67] for DT, DTD, and DTD-ML selection, respectively. There were significant differences between DT selection and the other methods. Because the task was to interact with a dictionary-like interface without any temporal limitation, and because the dwell interface did not require physical activity, the scores for the mental, physical, and temporal demand were smaller than the other scores. In terms of the performance and frustration scores, DT selection was inferior to DTD and DTD-ML selection, which is consistent with the quantitative results.

The averages and ranges of the overall SUS were 31.88 [22.5–40.0], 60.62 [47.5–70.0], and 69.38 [55.0–77.5] for DT, DTD, and DTD-ML selection, respectively. There were significant differences here, as well, between DT selection and the other selections. For all questions except Q6, "I thought there was too much inconsistency in this system," the scores for DT, DTD, and DTD-ML selection became higher in this order. Regarding the inconsistency, DTD selection had the highest score. DTD-ML selection achieved the best ratio; however, some intents to select were mistakenly detected as intent not to select. In other words, false negatives affected this result. For Q10, "I needed to learn a lot of things before I could get going with this system," there was no significant difference among the selection methods. Because we did not conduct a practice session for each method and the participants could interact with the interface by using each method, the scores became high with no significant differences. That is, the learning cost for dwell selection seems small regardless of the methods.

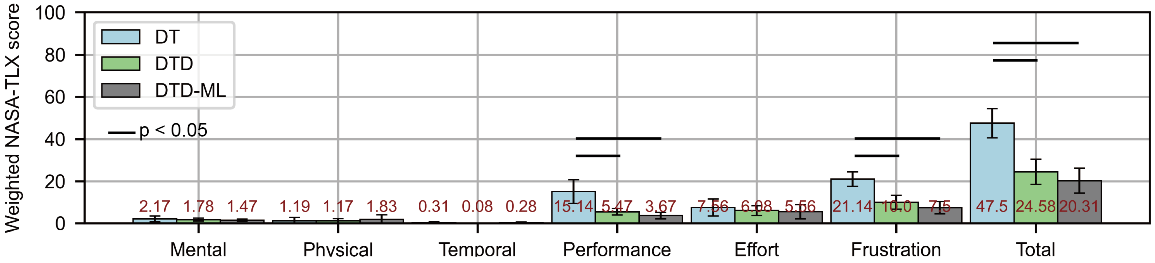


Fig. 8. NASA-TLX test results; lower values indicate better scores.

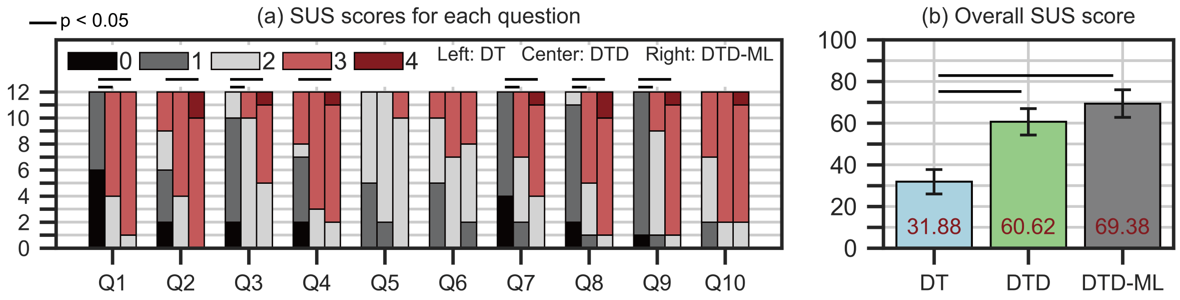


Fig. 9. SUS test results. (a) Bar chart showing the adjusted scores for each question in order of DT, DTD, and DTD-ML selection, where 0 (black) indicates the worst score and 4 (red) indicates the best score. (b) Box plot showing the overall scores (higher is better) among the participants. The specific questions are listed in the Supplementary.

6.4.3 Prediction Delays. We also measured the times required to create features and predict intent. The experimental PC was an Alienware Aurora R9 (CPU: Intel(R) Core^{TM} i9-9900 @ 3.10 GHz; RAM: 32.0 GB; OS: Windows 10 Version 21H2). The averages and ranges of the times for feature creation and for prediction were 3.55 ms [2.22–14.12] and 0.21 ms [0.12–1.37], respectively. The delay in comparison to DTD selection averaged 3.76 ms [2.35–14.60]. Because the eye-tracking frequency in Experiment 2 was 1,200 Hz (i.e., 0.83 ms/sample), the prediction could not finish within one sample. However, when using DTD-ML selection for interaction, such a delay does not seem critical.

7 LIMITATIONS AND FUTURE RESEARCH SPACE

7.1 Limitations on Applicable Interfaces and Interaction on DTD-ML

We showed that DTD-ML works for a dictionary-like interface whose contents have a size of at least 2.0° (2.3 cm in this experimental setting). We limited the target size to avoid issues related to eye-tracking accuracy. However, smaller sizes than 2.0° are used for tab-icons on the Windows 10 desktop and close buttons on a web browser. Still, a target size of 2.0° approximately reflects the desktop icons for the “medium icons” setting on Windows 10, which justifies our experimental setting. Moreover, some contents in image search on Microsoft Edge are often larger than 4.0° (approximately 4.5 cm in the experimental setting) with display zoom setting of 100%, and those contents are positioned in a grid layout with small margins between contents, which is similar to the interface used in Experiment 2. Thus, DTD-ML would work for a common interface design with a content size of at least 2.0°. Also, even though the *leave-one-task-out* result had an insufficient AUC value, the results of Experiment 2, whose interface and task differed from those of Experiment 1, show the robustness against the Midas-touch problem. However, the capability for selection of objects other than a character, word, sentence, or image is still unexplored, and thus further investigation is needed.

Moreover, the use of DTD-ML is limited to "selection." Other interactions such as activating a command and opening a menu are also necessary for more realistic use of gaze-based interaction. One solution using DTD-ML would be two-step manipulation, similar to right-clicking: a first selection would open a menu on a dwelled target, and a second selection would activate a command mapped to the dwelled menu item on the target. While this design is not new, because DTD-ML offers a robust trigger for opening a menu, it can prevent occlusion due to unwanted menu opening. This would also be useful for gaze-gesture research which uses dwell selection for trigger gesture detection (e.g., [1, 14, 31, 39, 70]). Another solution would be to combine with a second modality (e.g., [10, 54–56]). Note that the main contribution of this work is the establishment of an essential "selection" system like left-clicking a mouse, and thus these limitations should be explored.

7.2 Participant Dependency

We achieved strong prediction results for the participants because the features did not include user-dependent gaze data. Moreover, in Experiment 2, users whose gaze data had not been used for the ML model could use DTD-ML to select targets with similar effectiveness to users whose gaze data had been used. Because we did not use the original values for **pupil**, we eliminated the effect of pupil diameters. However, it is known that pupil diameters decrease with age [6], and further investigation will thus be needed to test our method with a very diverse range of users.

7.3 Application to DT Selection

By changing the threshold of the prediction probability, we can deal with the trade off between the robustness against the Midas-touch problem and the ease of selection. This is similar to research on tuning the dwell-time to prevent the Midas-touch problem and achieve fast selection, to which our work can contribute. For example, we could reduce the probability threshold for dwell-typing according to the probability that a key is typed. Moreover, the basic concept of DTD-ML detection is the same as that of DT detection, which uses only the dwell-time. Thus, we can also apply our method to the research on DT selection (e.g., [11, 76]) to improve the performance.

7.4 Use of Intent Prediction

We expect that our model is useful for other interactions. The most promising application is a gaze-supported system combining gaze and other modalities, which other researchers have attempted to develop as an AR/VR interface (e.g., [13, 43, 58]). In general, the intent not to select entails many aspects, such as paying attention or expressing intent, in terms of why the user looks at something. It would be difficult for our model to predict such varied intents because of its current limitations. However, advancement based on this research should lead to further use of gaze-based intent prediction and the development of real-world applications.

8 CONCLUSION

We have developed a dwell selection system that uses ML-based prediction of a user's intent to select a target. As features for the ML-based prediction, we used gaze movement, pupil diameter, and vergence, which is linked to a user's dwell action. To create the intent prediction model, we first conducted Experiment 1 on labeling of user intent with five tasks and then calculated features. The results showed that our model could predict a user's intent with a high AUC value of 0.903: specifically, 0.898 for prediction independent of the user and 0.880 for prediction independent of the eye-tracker. The results of Experiment 2 showed that DTD-ML selection could prevent 40.2% of unwanted selection in comparison to DTD selection, and that it had equal or better NASA-TLX and SUS scores than DT and DTD selection. Our approach of intent prediction should greatly contribute to system development for various interactive situations, and advancement based on our research should lead to further use of gaze-based intent prediction. Further exploration of parameters (i.e., the dwell-time, dispersion threshold, and window size) and investigation with various tasks will improve the intent prediction.

ACKNOWLEDGMENTS

This work is supported by JSPS, KAKENHI Grant Number 21J10301, Tateisi Science and Technology Foundation, and Marubun Research Promotion Foundation.

REFERENCES

[1] Sunggeun Ahn, Stephanie Santosa, Mark Parent, Daniel Wigdor, Tovi Grossman, and Marcello Giordano. 2021. StickyPie: A Gaze-Based, Scale-Invariant Marking Menu Optimized for AR/VR. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 739, 16 pages. https://doi.org/10.1145/3411764.3445297

[2] Sunggeun Ahn, Jeongmin Son, Sangyoon Lee, and Geehyuk Lee. 2020. Verge-It: Gaze Interaction for a Binocular Head-Worn Display Using Modulated Disparity Vergence Eye Movement. In *Proceedings of the 2020 CHI Extended Abstracts on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 264:1–7. https://doi.org/10.1145/3334480.3382908

[3] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2623–2631. https://doi.org/10.1145/3292500.3330701

[4] Borji Ali and Itti Laurent. 2013. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013), 185–207.

[5] Roman Bednarik, Hana Vrzakova, and Michal Hradis. 2012. What Do You Want to Do Next: A Novel Approach for Intent Prediction in Gaze-based Interaction. In *Proceedings of the 2012 ACM Symposium on Eye Tracking Research & Applications* (Santa Barbara, California) *(ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 83–90. https://doi.org/10.1145/2168556.2168569

[6] James E. Birren, Roland C. Casperson, and Jack Botwinick. 1950. Age Changes in Pupil Size. *Journal of Gerontology* 5, 3 (07 1950), 216–221. https://doi.org/10.1093/geronj/5.3.216 arXiv:https://academic.oup.com/geronj/article-pdf/5/3/216/1647183/5-3-216.pdf

[7] Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 7 (1997), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[8] John Brooke. 1996. *Usability Evaluation in Industry.* CRC Press, Chapter SUS-A Quick and Dirty Usability Scale, 189–194.

[9] Çağla Çığ Karaman and Tevfik Metin Sezgin. 2018. Gaze-based predictive user interfaces: Visualizing user intentions in the presence of uncertainty. *International Journal of Human-Computer Studies* 111 (2018), 78–91. https://doi.org/10.1016/j.ijhcs.2017.11.005

[10] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/2818346.2820752

[11] Myungguen Choi, Daisuke Sakamoto, and Tetsuo Ono. 2020. Bubble Gaze Cursor + Bubble Gaze Lens: Applying Area Cursor Technique to Eye-Gaze Interface. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20)*. Association for Computing Machinery, New York, NY, USA, Article 11, 10 pages. https://doi.org/10.1145/3379155.3391322

[12] Manfred Clynes. 1962. The Non-Linear Biological Dynamics of Unidirectional Rate Sensitivity Illustrated by Analog Computer Analysis, Pupillary Reflex to Light and Sound, and Heart Rate Behavior. *Annals of the New York Academy of Sciences* 98, 4 (1962), 806–845. https://doi.org/10.1111/j.1749-6632.1962.tb30600.x arXiv:https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.1962.tb30600.x

[13] Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. 2021. Towards Gaze-Based Prediction of the Intent to Interact in Virtual Reality. In *ACM Symposium on Eye Tracking Research and Applications* (Virtual Event, Germany) *(ETRA '21 Short Papers)*. Association for Computing Machinery, New York, NY, USA, Article 2, 7 pages. https://doi.org/10.1145/3448018.3458008

[14] William Delamare, Teng Han, and Pourang Irani. 2017. Designing a Gaze Gesture Guiding System. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 26, 13 pages. https://doi.org/10.1145/3098279.3098561

[15] Heiko Drewes, Mohamed Khamis, and Florian Alt. 2018. Smooth Pursuit Target Speeds and Trajectories. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia* (Cairo, Egypt) *(MUM '18)*. Association for Computing Machinery, New York, NY, USA, 139–146. https://doi.org/10.1145/3282894.3282913

[16] Tobii Dynavox. 2021. Assistive technology for communication/AAC - Tobii Dynavox. https://www.tobiidynavox.com/ (Retrieved January 27, 2021).

[17] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 2015. Orbits: Gaze Interaction for Smart Watches using Smooth Pursuit Eye Movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 457–466. https://doi.org/10.1145/2807442.2807499

[18] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 961–970.

[19] Anna Maria Feit, Lukas Vordemann, Seonwook Park, Caterina Berube, and Otmar Hilliges. 2020. Detecting Relevance during Decision-Making from Eye Movements for UI Adaptation. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Full Papers)*. Association for Computing Machinery, New York, NY, USA, Article 10, 11 pages. https://doi.org/10.1145/3379155.3391321

[20] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 2017. Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1118–1130. https://doi.org/10.1145/3025453.3025599

[21] Pedro Figueiredo and Manuel J. Fonseca. 2018. EyeLinks: A Gaze-Only Click Alternative for Heterogeneous Clickables. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 307–314. https://doi.org/10.1145/3242969.3243021

[22] Christoph Gebhardt, Brian Hecox, Bas van Opheusden, Daniel Wigdor, James Hillis, Otmar Hilliges, and Hrvoje Benko. 2019. Learning Cooperative Personalized Policies from Gaze Data. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 197–208. https://doi.org/10.1145/3332165.3347933

[23] John Hansen, Anders Johansen, Dan Hansen, Kenji Ito, and Satoru Mashino. 2003. Command Without a Click: Dwell Time Typing by Mouse and Gaze Selections. In *Proceedings of Human-Computer Interaction (INTERACTA '03)*. IFIP, 121–128.

[24] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[25] Mary Hayhoe and Dana Ballard. 2005. Eye Movements in Natural Behavior. *Trends in Cognitive Sciences* 9, 4 (2005), 188–194. https://doi.org/10.1016/j.tics.2005.02.009

[26] Eckhard H. Hess and James M. Polt. 1960. Pupil Size as Related to Interest Value of Visual Stimuli. *Science* 132 (1960), 349–350.

[27] Anthony Hornof, Anna Cavender, and Rob Hoselton. 2004. EyeDraw: A System for Drawing Pictures with the Eyes. In *Proceedings of the 2004 CHI Extended Abstracts on Human Factors in Computing Systems* (Vienna, Austria) *(CHI EA '04)*. Associati0on for Computing Machinery, New York, NY, USA, 1251–1254. https://doi.org/10.1145/985921.986036

[28] Anthony J. Hornof and Anna Cavender. 2005. EyeDraw: Enabling Children with Severe Motor Impairments to Draw with Their Eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 161–170. https://doi.org/10.1145/1054972.1054995

[29] Thomas E. Hutchinson, K. Preston White, Worthy N. Martin, Kelly C. Reichert, and Lisa A. Frey. 1989. Human-computer Interaction using Eye-gaze Input. *IEEE Transactions on Systems, Man, and Cybernetics* 19, 6 (1989), 1527–1534. https://doi.org/10.1109/21.44068

[30] Toshiya Isomoto, Toshiyuki Ando, Buntarou Shizuki, and Shin Takahashi. 2018. Dwell Time Reduction Technique Using Fitts' Law for Gaze-Based Target Acquisition. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (Warsaw, Poland) *(ETRA '18)*. Association for Computing Machinery, New York, NY, USA, 26:1–26:7. https://doi.org/10.1145/3204493.3204532

[31] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. 2020. Gaze-based Command Activation Technique Robust Against Unintentional Activation using Dwell-then-Gesture. In *Proceedings of Graphics Interface 2020* (University of Toronto) *(GI '20)*. Canadian Human-Computer Communications Society / Société canadienne du dialogue humain-machine, 256–266. https://doi.org/10.20380/GI2020.26

[32] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. 2021. Relationship between Dwell-Time and Model Human Processor for Dwell-based Image Selection. In *Proceedings of the 2021 ACM Symposium on Applied Perception* (virtual) *(SAP '21)*. Association for Computing Machinery, Article 6, 5 pages. https://doi.org/10.1145/3474451.3476240

[33] Robert J. K. Jacob. 1990. What You Look at is What You Get: Eye Movement-based Interaction Techniques. In *Proceedings of the 1990 CHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/97243.97246

[34] Robert J. K. Jacob. 1991. The Use of Eye Movements in Human-computer Interaction Techniques: What You Look at is What You Get. *ACM Transaction on Information Systems* 9, 2 (1991), 152–169.

[35] Robert. J. K. Jacob. 1993. Eye Movement-Based Human-Computer Interaction Techniques: Toward Non-Command Interfaces. *Advances in Human-Computer Interaction* 4 (1993), 151–190.

[36] Joaquin Jadue, Gino Slanzi, Luis Salas, and Juan D. Velásquez. 2015. Web User Click Intention Prediction by Using Pupil Dilation Analysis. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '15, Vol. 1)*. IEEE /

Association for Computing Machinery, New York, NY, USA, 433–436. https://doi.org/10.1109/WI-IAT.2015.221

[37] Dagmar Kern, Paul Marshall, and Albrecht Schmidt. 2010. Gazemarks: Gaze-Based Visual Placeholders to Ease Attention Switching. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2093–2102. https://doi.org/10.1145/1753326.1753646

[38] Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. 2016. GazeTouchPass: Multimodal Authentication Using Gaze and Touch on Mobile Devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 2156–2164. https://doi.org/10.1145/2851581.2892314

[39] Taejun Kim, Auejin Ham, Sunggeun Ahn, and Geehyuk Lee. 2022. Lattice Menu: A Low-Error Gaze-Based Marking Menu Utilizing Target-Assisted Gaze Gestures on a Lattice of Visual Anchors. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 277, 12 pages. https://doi.org/10.1145/3491102.3501977

[40] Dominik Kirst and Andreas Bulling. 2016. On the Verge: Voluntary Convergences for Accurate and Precise Timing of Gaze Input. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1519–1525. https://doi.org/10.1145/2851581.2892307

[41] Shinya Kudo, Hiroyuki Okabe, Taku Hachisu, Michi Sato, Shogo Fukushima, and Hiroyuki Kajimoto. 2013. Input Method Using Divergence Eye Movement. In *Proceedings of the 2013 CHI Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 1335–1340. https://doi.org/10.1145/2468356.2468594

[42] Michael F. Land and Mary Hayhoe. 2001. In what ways do eye movements contribute to everyday activities? *Vision Research* 41, 25 (2001), 3559–3565. https://doi.org/10.1016/S0042-6989(01)00102-X

[43] Feiyu Lu, Shakiba Davari, and Doug Bowman. 2021. Exploration of Techniques for Rapid Activation of Glanceable Information in Head-Worn Augmented Reality. In *Symposium on Spatial User Interaction* (Virtual Event, USA) *(SUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 14, 11 pages. https://doi.org/10.1145/3485279.3485286

[44] Christof Lutteroth, Moiz Penkar, and Gerald Weber. 2015. Gaze vs. Mouse: A Fast and Accurate Gaze-Only Click Alternative. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 385–394. https://doi.org/10.1145/2807442.2807461

[45] Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. 2009. Fast Gaze Typing with an Adjustable Dwell Time. In *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 357–360. https://doi.org/10.1145/1518701.1518758

[46] Päivi Majaranta, Anne Aula, and Kari-Jouko Räihä. 2004. Effects of Feedback on Eye Typing with a Short Dwell Time. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications* (San Antonio, Texas) *(ETRA '04)*. Association for Computing Machinery, New York, NY, USA, 139–146. https://doi.org/10.1145/968363.968390

[47] Päivi Majaranta, I. Scott MacKenzie, Anne Aula, and Kari-Jouko Räihä. 2006. Effects of Feedback and Dwell Time on Eye Typing Speed and Accuracy. *Universal Access in the Information Society* 5, 2 (2006), 199–208. https://doi.org/10.1007/s10209-006-0034-z

[48] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376479

[49] Pallavi Mohan, Wooi Boon Goh, Chi-Wing Fu, and Sai-Kit Yeung. 2018. DualGaze: Addressing the Midas Touch Problem in Gaze Mediated VR Interaction. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct '18)*. 79–84. https://doi.org/10.1109/ISMAR-Adjunct.2018.00039

[50] Martez E. Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. 2017. Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2558–2570. https://doi.org/10.1145/3025453.3025517

[51] Aanand Nayyar, Utkarsh Dwivedi, Karan Ahuja, Nitendra Rajput, Seema Nagar, and Kuntal Dey. 2017. OptiDwell: Intelligent Adjustment of Dwell Click Time. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 193–204. https://doi.org/10.1145/3025171.3025202

[52] Abdul Moiz Penkar, Christof Lutteroth, and Gerald Weber. 2013. Eyes Only: Navigating Hypertext with Gaze. In *14th IFIP TC 13 International Conference on Human-Computer Interaction – INTERACT 2013*. Springer Berlin Heidelberg, Berlin, Heidelberg, 153–169.

[53] Ken Pfeuffer, Yasmeen Abdrabou, Augusto Esteves, Radiah Rivu, Yomna Abdelrahman, Stefanie Meitner, Amr Saadi, and Florian Alt. 2021. ARtention: A Design Space for Gaze-adaptive User Interfaces in Augmented Reality. *Computers & Graphics* 95 (2021), 1–12. https://doi.org/10.1016/j.cag.2021.01.001

[54] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, and Hans Gellersen. 2014. Gaze-touch: Combining Gaze with Multi-touch for Interaction on the Same Surface. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 509–518. https://doi.org/10.1145/2642918.2647397

[55] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, Yanxia Zhang, and Hans Gellersen. 2015. Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte,

NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 373–383. https://doi.org/10.1145/2807442.2807460

[56] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. 2017. Gaze + Pinch Interaction in Virtual Reality. In *Proceedings of the 5th Symposium on Spatial User Interaction* (Brighton, United Kingdom) *(SUI '17)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/3131277.3132180

[57] Jimin. Pi and Bertram. E. Shi. 2017. Probabilistic Ajustment of Dwell Time for Eye Typing. In *10th International Conference on Human System Interactions (HSI)*. IEEE, 251–257.

[58] Robin Piening, Ken Pfeuffer, Augusto Esteves, Tim Mittermeier, Sarah Prange, Philippe Schröder, and Florian Alt. 2021. Looking for Info: Evaluation of Gaze Based Information Retrieval in Augmented Reality. In *18th IFIP TC 13 International Conference on Human-Computer Interaction – INTERACT 2021*. Springer International Publishing, 544–565.

[59] Panwar Prateek, Sarcar Sayan, and Samanta Debasis. 2012. EyeBoard: A Fast and Accurate Eye Gaze-Based Text Entry System. In *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI '12)*. 1–8. https://doi.org/10.1109/IHCI.2012.6481793

[60] Kari-Jouko Räihä and Saila Ovaska. 2012. An Exploratory Study of Eye Typing Fundamentals: Dwell Time, Text Entry Rate, Errors, and Workload. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 3001–3010. https://doi.org/10.1145/2207676.2208711

[61] Dario D. Salvucci and John R. Anderson. 2000. Intelligent Gaze-Added Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 273–280. https://doi.org/10.1145/332040.332444

[62] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Gardens, Florida, USA) *(ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. https://doi.org/10.1145/355017.355028

[63] Simon Schenk, Marc Dreiser, Gerhard Rigoll, and Michael Dorr. 2017. GazeEverywhere: Enabling Gaze-only User Interaction on an Unmodified Desktop PC in Everyday Scenarios. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3034–3044. https://doi.org/10.1145/3025453.3025455

[64] Asma Shakil, Christof Lutteroth, and Gerald Weber. 2019. CodeGazer: Making Code Navigation Easy and Natural With Gaze Input. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300306

[65] Linda E. Sibert and Robert J. K. Jacob. 2000. Evaluation of Eye Gaze Interaction. In *Proceedings of the 2000 CHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 281–288. https://doi.org/10.1145/332040.332445

[66] Henrik Skovsgaard, Kari-Jouko Räihä, and Martin Tall. 2011. Computer Control by Gaze. In *Gaze Interaction and Aplications of Eye Tracking: Advances in Assistive Technologies*, Päivi Majaranta, Hirotaka Aoki, Mick Donegan, Witzner Hansen Dan, John Paulin Hansen, Aulikki Hyrskykari, and Kari-Jouko Räihä (Eds.). IGI Global, Hershey, PA, Chapter 9, 78–103.

[67] Sophie Stellmach and Raimund Dachselt. 2012. Look & Touch: Gaze-supported Target Acquisition. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2981–2990. https://doi.org/10.1145/2207676.2208709

[68] Geoffrey Tien and M. Stella Atkins. 2008. Improving Hands-Free Menu Selection Using Eyegaze Glances and Fixations. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (Savannah, Georgia) *(ETRA '08)*. Association for Computing Machinery, New York, NY, USA, 47–50. https://doi.org/10.1145/1344471.1344482

[69] Jayson Turner, Jason Alexander, Andreas Bulling, and Hans Gellersen. 2015. Gaze+RST: Integrating Gaze and Multitouch for Remote Rotate-Scale-Translate Tasks. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 4179–4188. https://doi.org/10.1145/2702123.2702355

[70] Mario H. Urbina, Maike Lorenz, and Anke Huckauf. 2010. Pies with EYEs: The Limits of Hierarchical Pie Menus in Gaze Control. In *Proceedings of the 2010 ACM Symposium on Eye-Tracking Research & Applications* (Austin, Texas) *(ETRA '10)*. Association for Computing Machinery, New York, NY, USA, 93–96. https://doi.org/10.1145/1743666.1743689

[71] Eduardo Velloso, Flavio Luiz Coutinho, Andrew Kurauchi, and Carlos H Morimoto. 2018. Circular Orbits Detection for Gaze Interaction Using 2D Correlation and Profile Matching Algorithms. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (Warsaw, Poland) *(ETRA '18)*. Association for Computing Machinery, New York, NY, USA, Article 25, 9 pages. https://doi.org/10.1145/3204493.3204524

[72] Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 2013. Pursuits: Spontaneous Interaction with Displays Based on Smooth Pursuit Eye Movement and Moving Targets. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 439–448. https://doi.org/10.1145/2493432.2493477

[73] Oleg Špakov, Poika Isokoski, Jari Kangas, Deepak Akkil, and Päivi Majaranta. 2016. PursuitAdjuster: An Exploration into the Design Space of Smooth Pursuit-based Widgets. In *Proceedings of the 2016 ACM Symposium on Eye Tracking Research & Applications* (Charleston, South

Carolina) *(ETRA '16)*. Association for Computing Machinery, New York, NY, USA, 287–290. https://doi.org/10.1145/2857491.2857526

[74] Colin Ware and Harutune H. Mikaelian. 1987. An Evaluation of an Eye Tracker As a Device for Computer Input. In *Proceedings of the 1987 CHI/GI Conference on Human Factors in Computing Systems and Graphics Interface* (Toronto, Ontario, Canada) *(CHI '87)*. Association for Computing Machinery, New York, NY, USA, 183–188. https://doi.org/10.1145/29933.275627

[75] Pingmei Xu, Yusuke Sugano, and Andreas Bulling. 2016. *Spatio-Temporal Modeling and Prediction of Visual Attention in Graphical User Interfaces*. Association for Computing Machinery, New York, NY, USA, 3299–3310. https://doi.org/10.1145/2858036.2858479

[76] Xinyong Zhang, Xiangshi Ren, and Hongbin Zha. 2008. Improving Eye Cursor's Stability for Eye Pointing Tasks. In *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 525–534. https://doi.org/10.1145/1357054.1357139

[77] Xinyong Zhang, Pianpian Xu, Qing Zhang, and Hongbin Zha. 2011. Speed-Accuracy Trade-off in Dwell-Based Eye Pointing Tasks at Different Cognitive Levels. In *Proceedings of the 1st International Workshop on Pervasive Eye Tracking & Mobile Eye-Based Interaction* (Beijing, China) *(PETMEI '11)*. Association for Computing Machinery, New York, NY, USA, 37–42. https://doi.org/10.1145/2029956.2029967