University of Tsukuba

Doctoral Thesis

Study on Eye Behavior-Based Intent Detection for Dwell Selection

Toshiya Isomoto
Graduate School of Science and Technology
Degree Programs in Systems and Information Engineering

A thesis submitted for the degree of
Doctor of Philosophy in Engineering
September 2023

Study on Eye Behavior-Based Intent Detection for Dwell Selection
Toshiya Isomtoo

Graduate School of Science and Technology
Degree Programs in Systems and Information Engineering
UNIVERSITY OF TSUKUBA
A thesis submitted for the *Doctor of Philosophy in Engineering.* September, 2023

Abstract

While developing a new interaction method, it is crucial to explore how a computer detects user intent to interact with the computer. In particular, gaze-based interaction, which utilizes human natural eye behavior such as fixation, saccade, vergence, smooth pursuit, and pupil diameter, has been the primary focus of the researcher for exploring user intent detection methods. Human eyes have some common functions, such as subconsciously observing visual information and showing our attention and intent through the eye. Therefore, it is possible to detect user attention and intent by observing changes in eye behaviors without requiring additional behavior for users. This implicitness of eye behavior is attractive for interaction, and hence gaze-based interaction, which utilizes eye behavior, has been widely researched. We developed the user intent detection methods for dwell selection, which is a gaze-based interaction.

Dwell selection method utilizes the human eye behavior of "looking." For dwell selection on the object that users want to select, users are required to find the object and keep looking at it. More systematically, in a 2D display, dwell selection is triggered when the x and y gaze coordinates on display are inside a graphical user interface (GUI) object for a certain duration, referred to as *dwell time*. Dwell time is an indispensable parameter for detecting user intent to select a GUI, and gaze coordinates are an indispensable parameter for determining which GUIs are desirable to the user. Although utilizing "looking" without any additional action is attractive from the aspect of implicit use of eye behavior, it may result in a mis-detection of user intent. The mis-detection causes unwanted selection, referred to as Midas-touch, and solving Midas-touch has been the goal of dwell selection research. However, it has been 30 years since dwell selection was first researched, and the solution has not been derived yet. Furthermore, solving Midas-touch with intent detection using only dwell time seems challenging.

In this thesis, we show user intent detection models to extend the determination of dwell time and the method itself. One model extends the current determination

of dwell time by incorporating natural eye behavior and human decision-making processes. The current determination is based on the optimization of the speed and accuracy of dwell selection. Although dwell time plays a significant role in user intent detection for dwell selection, the determination method of dwell time lacks incorporation of the human decision-making process. Another model extends user intent detection by incorporating multiple natural eye behaviors. While natural eye behaviors potentially reflect user intent, because they generally rely on users, ambient environment, and interaction situations, identifying the eye behaviors and characteristics that are useful for interpreting user intent is not simple. Therefore, to interpret the user intent from such eye behaviors, we adopted a machine learning (ML) based method and developed an ML model that can interpret the intent using eye behavior features. Lastly, we demonstrate the use of our models for dwell selection and how our model extends the gaze-based interaction.

Acknowledgements

Behind this thesis lies the support and help of numerous people to whom I wish to express my gratitude. First, I would like to express my sincere thanks to my advisors Prof. Buntarou Shizuki, Assoc. Prof. Shin Takahashi, Asst. Prof. Ikkaku Kawaguchi, and Simona Vasilache for their continuous support of my undergraduate, graduate, and doctoral studies, and for their patience and motivation. Their guidance helped me grow into the researcher that I am today through the completion of this thesis. I could not have had better advisors and mentors for my Ph.D. journey.

I want to thank my labmates —both old and new— for the long days of fun and conversation about life and research. I will fondly look back at the last seven years of moments we have had together. I feel lucky to count you all among my friends who have been there and supported me day after day.

My sincere thanks also go to Ryoichi Miyazaki, who introduced me to the world of academia while I was a Kosen student and encouraged me to apply to the University of Tsukuba. Without his encouragement and instruction, I would not have pursued my research career. I also thank Dr. Shota Yamanaka at Yahoo! Japan Corporation for the opportunity to intern with his team. My time with you was enormously inspiring, and it was a fantastic experience to spend four years with you. Also, I thank the Japan Society for the Promotion of Science and the Tateisi Science and Technology Foundation for funding my research projects.

I could not have reached this point without the support and love of my family. I would like to thank my parents and brother, who have continuously supported me and encouraged me to start this journey. Finally, I would like to thank all the friends I have made in Yamaguchi and Tsukuba, and all my friends back home, who I have made this journey so special by sharing it with me.

PUBLICATIONS

This work has been published in peer-reviewed publications at conferences and journals. Below is the reference for these publications.

Chapter 3 Toshiya Isomoto, Shota Yamanaka, Buntarou Shizuki. Exploring Dwell Time from Human Cognitive Processes for Dwell Selection. Proceedings of the ACM on Human-Computer Interaction Volume 7, Issue ETRA, Article 159, pp.1-15. Association for Computing Machinery, New York, NY, USA, May 2023. DOI: 10.1145/3591128

Chapter 4 Toshiya Isomoto, Shota Yamanaka, Buntarou Shizuki. Dwell Selection with ML-based Intent Prediction Using Only Gaze Data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), Vol.6, No.3, Article 120, pp.1-21. Association for Computing Machinery, New York, NY, USA, May 2023, September 2022. DOI: 10.1145/3550301

Contents

ABSTRACT **iii**

ACKNOWLEDGEMENTS **v**

PUBLICATIONS **vii**

1 INTRODUCTION **1**
- 1.1 User Intent Detection For Interaction . . . 1
- 1.2 Gaze-Based Interaction . . . 2
- 1.3 Dwell Selection and Issues . . . 5
- 1.4 Research Questions . . . 7
- 1.5 Methodology . . . 7
- 1.6 Contributions . . . 9
- 1.7 Thesis Structure . . . 9

2 RELATED WORK **11**
- 2.1 Eye Behaviors for Gaze-Based Interaction . . . 11
- 2.2 Gaze-Based Interaction . . . 14
  - 2.2.1 Implicit Gaze Interaction . . . 14
  - 2.2.2 Explicit Gaze Interaction . . . 15
    - Interaction with Confirmation Button . . . 15
    - Interaction with Smooth Pursuit . . . 16
    - Interaction with Eye-Gesture . . . 16
  - 2.2.3 Multimodal Gaze Interaction . . . 18
- 2.3 Dwell Selection Challenges . . . 19
  - 2.3.1 Dwell Time for Dwell Selection . . . 19
  - 2.3.2 Midas-touch and Solution . . . 20
- 2.4 Position of This Thesis . . . 20

3 DETERMINATION OF DWELL TIME THROUGH FIXATION AND MHP **23**
- 3.1 Human Decision-Making Process via Model Human Processor . . . 24

3.1.1 Selection Tasks in MHP 25
3.1.2 Time Required for Competing Tasks 26
3.1.3 Relation Between Dwell Time and MHP 27
3.1.4 Models of Human Cognition and Behavior for Visual Search Tasks 29
3.2 Hypotheses 30
3.3 Experiment 31
3.3.1 Participants and Apparatus 31
3.3.2 Selection Method 31
3.3.3 Selection Task and Interface 32
Simple Task 34
Key Task 34
Icon Task 34
Word Task 35
Image Task 35
3.3.4 Results 36
Number of Fixations 37
Duration of Fixation 37
3.4 Model Deriving Dwell Time 39
3.4.1 Equations of Model 40
3.4.2 Slope in Our Model 41
3.4.3 Minimum $D_{fixation}$ for Each Task 41
3.4.4 Range of $D_{fixation}$ 42
3.5 Applying Our Model for Dwell Selection 43
3.6 Conclusions 44

4 USER INTENT DETECTION WITH MULTIPLE NATURAL EYE BEHAVIORS FOR DWELL SELECTION 47
4.1 Our Dwell Selections 48
4.1.1 Dwell Time and Dispersion (DTD) Based User Intent Detection (DTD Detection) 50
4.1.2 Intent Detection with an ML Model 51
4.1.3 Target Selection 51
4.2 Experiment 1: Labeling of User Intent 51
4.2.1 Participants and Apparatus 51
4.2.2 Tasks 52
Letter Task 53
Word Task 53
Sentence Task 54
Image Task 54

Movie Task . . . 54
Intent labeling . . . 55
4.2.3 Procedure . . . 55
4.2.4 Labeling Results . . . 55
4.3 Model Detecting User Intent for Dwell Selection . . . 56
4.3.1 Data Processing . . . 56
4.3.2 Features . . . 57
4.3.3 Metrics for Evaluation . . . 59
4.3.4 Creating ML Model . . . 59
Overall detection Results . . . 60
Detection Results for Participants . . . 60
Detection Results for Tasks . . . 60
Detection Results for Frequencies . . . 61
Detection Results for Window Size . . . 61
Detection Results for Other Classifiers . . . 61
Use of Task- and Participant-dependent Gaze Data . . . 61
Feature Importance . . . 62
4.4 Experiment 2: Performance Evaluation . . . 62
4.4.1 Participants and Apparatus . . . 62
4.4.2 Task . . . 63
4.4.3 Procedure . . . 64
4.4.4 Results . . . 64
Quantitative Results . . . 64
Qualitative Results . . . 65
Detection Delays . . . 66
4.5 Discussion . . . 67
4.5.1 Limitations on Applicable Interfaces and Interaction on DTD-ML . . . 67
4.5.2 Participant Dependency . . . 68
4.5.3 Application to DT Selection . . . 68
4.5.4 Exploring Parameters . . . 69
4.5.5 Feature Exploration . . . 69
4.6 Conclusion . . . 70

5 CONCLUSIONS **73**
5.1 Use of Our Dwell Time Determination Model . . . 73
5.2 Use of Our Intent Detection Model . . . 74
5.3 Conclusions . . . 75

List of Figures

1.1 Overview of our work. 8

2.1 Eye behavior used for gaze-based interaction. 12

3.1 Overview of MHP. Image from Card et al. [CNM83]. 24
3.2 Participants' preferred dwell time for image selection task in our previous work [IYS21]. 27
3.3 Button press time for image selection task in our previous work [IYS21]. 28
3.4 Experimental environment. 31
3.5 Displays for five selection tasks. 33
3.6 Fixations we used (a and b) and did not use for the subsequent analysis (c and d). 36
3.7 D_{fixation} of the last fixation for each task and each N_{fixation}. For example, for image task, D_{fixation} averaged in 330 ms for N_{fixation} was six (i.e., the participants required six times of fixations to complete the trial. 38
3.8 D_{fixation} of last fixation against the sum of D_{fixation}s before last fixation. 38
3.9 Regression results with average D_{fixation} and linear equation (Equation 3.2) for each task. Gray plots are average D_{fixation} for each participant. Red plots are average for each N_{fixation}. 40
3.10 Regression results with average D_{fixation} and logarithmic equation (Equation 3.3) for each task. Gray plots are average D_{fixation} for each participant. Red plots are average for each N_{fixation}. 40

4.1 Overview of the DTD-ML selection system. 48
4.2 The display used for investigating dispersion in the preliminary experiment to determine the dispersion threshold. Points were instructed points where participants looked. 49
4.3 Results of the preliminary experiment to determine dispersion threshold: dispersion results for each dwell time (a) and position (b). . . . 49
4.4 Displays used for tasks. 52
4.5 Example of raw data for (a) **pupil** and (b) its downsampled values. 57

4.6 Feature importance for our intent detection method. 62
4.7 Interface used in Experiment 2. 63
4.8 Quantitative results for the selection of known targets. Values in parentheses indicate numbers of unwanted selections and total selections. 64
4.9 Quantitative results for the selection of search targets. Values in parentheses indicate number of unwanted selections and total selections. 65
4.10 NASA-TLX test results; lower values indicate better scores. 66
4.11 SUS test results. (a) Bar chart showing 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 overall scores (higher is better) among participants. 67

List of Tables

3.1 Tasks and their instructions described in [CNM83]. 25
3.2 Tasks and their required processes described in [CNM83]. 25
3.3 Tasks and their difficulties. “Known candidates” means whether or not a participant knew which keys/icons/words/images were shown in candidates before a task began. We assigned “difficulties” in accordance with the row “Similar task in MHP,” where each task requires a different number of required decision-making processes. The difference in “Known candidates” between the key and icon tasks results in a different difficulty even though the task in the MHP is the same. 32
3.4 N_{fixation} required for completing each task. The number in the parentheses is that of the participants. For example, twelve participants required two fixations in 20 trials to complete the key task. 37
3.5 Regression results for each task on a linear model of Equation 3.2: $a + b \times (N_{\text{fixation}}\text{-}1)$. 39
3.6 Regression results for each task on a logarithmic model of Equation 3.3: $a + b \times \log_2(N_{\text{fixation}})$. 39
3.7 Relation between slope and estimated slope using MHP. Units of all digits are in milliseconds. The estimated slope was calculated with the number of required processes *recognizing*, *classifying*, *matching* for each task. For example, in image task, four processes of *recognizing*, *classifying*, *matching*, and *requesting*, require $4\tau_c$. Because a *request* process can be regarded as the process of decision-making for tasks described in [CNM83], we exclude τ_c for *request* from one cognitive cycle, the estimated slope with MHP is similar to $3\tau_c$= 210 ms. 42
3.8 Summary of regression results for each task. 43

4.1 Guidelines for intent labeling. 55
4.2 Numbers of labeled classes. 56
4.3 Calculated features. In total, we used 127 ($= 80 + 35 + 12$) features for ML. 58

4.4 Summary of our intent detection. Values except for *all* are average values. MCC means Matthews correlation coefficient. We highlighted important results with aspects of contribution (red) and limitation (blue). 59

Chapter 1

INTRODUCTION

This thesis introduces eye behavior-based user intent detection methods and their use in dwell selection. Eye behavior is a natural human behavior that allows humans to subconsciously obtain visual information and show their attention and intent. Interaction with eye behavior is referred to as gaze-based interaction. Previous research on gaze-based interaction primarily focused on solving unwanted interactions caused by misdetection of user intent for interaction using only eye behavior. To solve this unwanted interaction, we base our work on previous human-computer interaction (HCI) research wherein the human natural eye behaviors were used as an input modality. In particular, we focused on dwell selection, a fundamental interaction method of "selection" in gaze-based interaction, presented by Jacob [Jac90]. In dwell selection, unwanted interaction, i.e., unwanted selection, in this case, is observed, and several efforts have been made to establish an ideal solution. Furthermore, we review previous work on gaze-based interaction, describe dwell selection and its issues and why we focus on dwell selection, investigate the relation between eye behaviors and user intent for interaction, and present two models based on the relation to address the issues.

1.1 User Intent Detection For Interaction

While developing an interaction method, the accurate detection of user intent is important to achieve accurate interactions. Various human body parts, such as the hand, head, foot, and mouth, serve as modalities for detecting user intent. As an example of mouse-based interaction, which is currently one of the most established interaction methods, users can express their intent to interact or not interact with a computer through actions like "left-click," "right-click," and "scroll" in a mouse-based interaction. Touch-based interaction offers another example, where a simple "tap" with fingers signifies the user's intent to interact with a smartphone. Moreover, hand gestures and voice are also used as modalities to detect user intent. Using a mouse or hand, users can manipulate a graphical user interface (GUI) by searching an object, landing a cursor, or moving a finger on the object, and then performing a "left clicking" or "tap"; the selection is triggered on the object. All those modalities enable users to explicitly show their intent. This capability arises from humans' dexterity in moving their hands and fingers with precision, enabling them to perform distinct gestures and vocalize specific words crucial for detecting user intent.

In recent years, interfaces have been designed to offer information linked to the user's surrounding environment, aligned with the concept of establishing ubiquitous computing. Ubiquitous computing envisions "a new way of thinking about computers in the world, one that takes into account the natural human environment and allows the computers themselves to vanish into the background" [Wei91]. Lots of researchers have put their effort into establishing ubiquitous computing as the future world. An example of such an interface in ubiquitous computing is the adaptive interface wherein a user's gaze upon a food, item triggers the presentation of additional information, such as the calorie count and allergen details. This adaptation with the user intent occurs seamlessly, requiring no further explicit actions. To make the computers in the background, the development of an intent detection method based on users' subconscious and natural behavior has garnered attention. Especially the multifaceted roles of human eyes in gathering information about the environment and signaling attention to others.

Furthermore, regarding the aspect of user intent detection based on human natural behavior, the exploration into the brain's electrical activity is also researched. The interaction utilizing the brain's electrical activity is known as brain-computer interaction (BCI), which operates sans physical movement of body parts, leveraging analysis of the brain's electrical signals to discern user intent. Since bodily movements are directed by signals originating in the brain, utilizing the brain's electrical activity can expand the interaction space beyond those dependent natural eye behavior. However, there are existing limitations in the approaches to sensing the brain signal (e.g., usability, technical challenges, and ethical challenge). In this context, our focus centers on utilizing the human eyes, given their proximity compared to the other body parts, as a means of reflecting subconscious human intent. Further rationale behind our emphasis on gaze-based interaction is elaborated upon in the next section.

1.2 Gaze-Based Interaction

Human eyes perform specific functions in our daily life, such as subconsciously observing visual information and showing our attention and intent through the eye. Based on the proverb "The eyes say more than the mouth," we can possibly detect the attention and intent of others by observing the subtle/unsubtle changes in their eye behaviors. Such eye behaviors could be a powerful modality for interaction if a computer could detect user intent to interact with itself through eye behaviors. The interaction using such eye behavior has been researched as gaze-based interaction. In this thesis, the word "gaze" is defined as the direction in which users look, and gaze-based interaction uses "gaze," which is detected through an eye tracker.

The interaction method based on eye behaviors, which are sampled through eye trackers, as a modality is called gaze-based interaction. In previous studies, "looking," "moving the eyes," and "blinking" have been used for user intent detection for gaze-based interaction. For example, users can manipulate a GUI by searching for an object and constantly "looking" at the object for a while; the selection is then triggered on the object. The functionality that the eyes move faster than other body parts, e.g., hands, feet, head, and mouth, is attractive for faster interaction. An interaction method that allows users to interact with a computer faster is the preferred method in the HCI field. Moreover, gaze-based interaction can be used as hands-free interaction. For accessibility, users with limited motor control, such as those who have amyotrophic lateral sclerosis (ALS), can interact with a computer using gaze-based interaction [Dyn21]. Such aspects of natural human eye behaviors have the potential to extend current interaction. Therefore, gaze-based interaction has been focused on the next interaction method following mouse- and touch-based interaction, which are the most established methods, as of 2023.

Eye-tracking technology has been developed for over 100 years and has been used to detect human visual attention. The pupil-center corneal reflection is one of the most commonly used techniques. The basic concept is to illuminate the eye and capture its image. The eye image is then used to identify the pupil center and reflection of the illuminators on the cornea. Further, image-processing algorithms are used to estimate a 3D model of the eyes and the position of the eye in space. Gaze is a direction that users look at, and it can be calculated using the reflection and pupil position. This 3D model also derives the user pupil position and diameter.

For developing an eye tracker, an improvement in its performance and a decrease in cost helped researchers further explore gaze-based interaction. While the roots of gaze-based interaction are in the 1980s [WM87, HWM+89], the eye-tracking system was not a widely adopted commercial equipment because it did not demonstrate sufficient eye-tracking performance in terms of frequency, accuracy, and precision. In the late 2010s, eye-tracking systems such as eye trackers by Tobii became more common commercial equipment for desktop computing. Currently, the eye-tracking system for head-mounted displays (HMD) (e.g., HTC Vive Pro Eye and HoloLens 2) has been developed and has become commercial equipment. Therefore, because researchers and developers can easily obtain precise eye behaviors, gaze-based interaction has attracted significant attention.

To the best of our knowledge, the first gaze-based interaction research was conducted by Colin and Mikaelian [WM87] in 1987. Numerous studies have been conducted to establish gaze-based interaction as a common interaction method to date. We categorize gaze-based interaction into three types: implicit, explicit, and multimodal gaze interactions. A detailed explanation is given in the following sections and Chapter 2.

Implicit Gaze Interaction

Generally, implicit interaction does not require any explicit action in addition to natural human behavior. User intent is detected using natural human behavior, and thus, the detection is implicitly done. Then, an interaction is triggered. Reliable intent detection is essential for implicit interaction, which has facilitated the research on intent detection based on natural human behavior.

Because most eye behaviors are subconsciously performed behaviors, their use for interaction is suitable to ensure implicit interaction. We categorize implicit interaction, specifically utilizing only natural human eye behavior, as implicit gaze interaction. Among various gaze-based interactions, the dwell selection, which utilizes the natural eye behavior of "looking," is the most relevant interaction method to implicit gaze interaction. A computer with dwell selection detects user intent from one natural eye behavior of "looking" at objects. Reliable intent detection is essential for implicit interaction, which has facilitated research on intent detection using natural human behavior. The implicit gaze interaction can deliver the potential of natural eye behaviors for good interaction. However, this has resulted in unwanted selection owing to the mis-detection of user intent from natural eye behaviors. Our goal is to establish a dwell selection that retains the benefits of interaction using natural eye behaviors while addressing the issue of the difficulty of user intent detection.

Explicit Gaze Interaction

In general, explicit interaction requires user-predefined actions, such as "clicking" with a mouse or moving hands or eyes in a specific manner. Such interactions are currently the mainstream interaction. Actions such as moving hands, eyes, head, or the whole body in a specific manner are referred to as gestures. The most important advantage of using explicit actions is the ease of accurate user intent detection because actions used for an interaction are designed to differentiate natural human behavior.

We categorize explicit interaction, specifically utilizing voluntary human eye behavior, as explicit gaze interaction. Most voluntary eye behaviors are adopted to be distinguishable from natural eye behaviors and realize the easiness of accurate user intent detection. Therefore, the probability of occurrence of mis-detection of user intent is less than implicit gaze interaction. Moreover, assigning commands to eye behaviors allows users to trigger the commands.

Multimodal Gaze Interaction

In contrast to the first two interactions that use eye behavior as a modality, multimodal gaze interaction employs eye behavior as an assistive modality alongside other modalities.

We categorize explicit interaction, specifically utilizing voluntary human eye behavior, as explicit gaze interaction. In the multimodal gaze interaction, eye behaviors are used as a cue to indicate user attention, whereas other modalities are used as a cue to indicate user intent to interact. One example is the look-and-touch principle, wherein a touch interaction is triggered where users look [SD12b]. Multimodal gaze interactions incorporate natural eye behaviors, which implicitly show user attention and ease of user intent detection of mouse, hand, and voice interaction.

1.3 Dwell Selection and Issues

The dwell selection method utilizes a human eye behavior of "looking," defined by Jacob [Jac90] as "if the user continues to look at the object for a sufficiently long time, it is selected without further operations." More systematically, in a 2D display, dwell selection is triggered when the x and y gaze coordinates on display are inside a GUI object for a certain duration called the dwell time. For dwell selection on the object that the users want to select, users are required to find the object and keep looking at the object. Ideally, the dwell selection does not require users to do any actions besides finding an object.

The user intent to select a GUI object is detected when measuring a duration that the gaze coordinates keep inside a GUI object over dwell time. For example, if we determine 1 s as the dwell time, a computer infers that a user wants to select the target that the user looks at when gaze coordinates keep being inside the target for over 1 s. Therefore, dwell time is an indispensable parameter for detecting user intent to select with a GUI, and gaze coordinates are an indispensable parameter for which GUIs are the user-desired GUI. Because dwell time roles detect user intent to select, researchers have explored the size of dwell time that should be used for developing a dwell selection and for comparing the performance of other interaction methods with dwell selection.

The goal of the research on dwell selection is to solve the long-time unsolved issue of **Midas-touch**, coined by Jacob [Jac90]. Its definition states, "Everywhere you look, something is activated; you cannot look anywhere without issuing a command." Therefore, Midas-touch is an issue where an object is accidentally selected. The cause of Midas-touch is mainly attributed to dwell time. In particular, a smaller dwell time may induce the mis-detection of user intent. For example, with the smallest dwell time (i.e., 0 ms), when gaze coordinates accidentally enter an object, the object is immediately selected, and hence, Midas-touch occurs. There is a possibility of correct detection of user intent. However, the detection becomes a mis-detection considering that most interaction with GUI requires users to search the GUI (i.e., looking at an object, understanding it, and deciding to select it), and may require a long duration. By using a larger dwell time, which is a simple solution, the mis-detection of user intent can be prevented; however, the time required for interaction becomes large, and even when using a large dwell time, if users continuously look at a target while thinking about something or observing the target, Midas-touch will occur. Therefore, researchers have explored a smaller dwell time that can prevent Midas-touch.

Because our eyes are constantly directed at something and moving, careful consideration of how we detect user intent from eye behavior is necessary to solve the Midas-touch. A majority of the methods involve adjusting dwell time according to the selection situations. For example, in dwell typing (i.e., dwell selection on a key), researchers used 180–600 ms as dwell times based on two perspectives: user preference and robustness against Midas-touch (e.g., [MAv09]). To select a key that is likely to be selected, using a small dwell time enables faster selection while using a large dwell time can prevent Midas-touch for a key that is unlikely to be selected. Hence, previous research on solving Midas-touch has explored a smaller dwell time that can prevent Midas-touch and allows faster interaction. Although it has been 30 years since dwell selection was developed, the solution has not been derived yet, and it looks difficult to solve Midas-touch with intent detection using only dwell time for solving Midas-touch.

Summary Intent detection for dwell selection is based on time-based (dwell time) and gaze coordinates. If the duration that the users continuously look at an object exceeds the dwell time, the selection is triggered on the object. An ideal dwell selection does not require additional voluntary eye behaviors and is suitable for implicit gaze interaction. However, dwell time-based user intent detection faces the issue of Midas-touch, which is an unwanted selection. Assuming we could solve Midas-touch using only natural human eye behavior without any additional voluntary eye behavior or voluntary behavior of other modalities, the potential of natural eye behaviors for gaze-based interaction, which is accessible for various users and fast interaction, is delivered. Therefore, we focus on developing a user intent detection method from the viewpoints of exploring dwell time and incorporating multiple natural eye behaviors.

1.4 Research Questions

This thesis aims to reveal how user intent to select or not select is detected by natural eye behaviors and establish dwell selection as a daily interaction method. To achieve this goal, we aim to develop a user intent detection method by utilizing natural human eye behaviors during the interaction. Concerning the aforementioned factors, we pose two research questions about eye behaviors and user intent in the context of dwell selection.

RQ1 *How should we determine dwell time?* While previous research explored dwell time by focusing on the speed and accuracy of dwell selection, we intend to determine dwell time by incorporating natural human eye behaviors and human decision-making processes. By revealing a relationship between natural human eye behaviors and human decision-making processes, we aim to demonstrate a new determination method of dwell time.

RQ2 *Can eye behavior reveal user intent to interact?* Dwell selection has relied on gaze coordinates and time-threshold-based user intent detection. However, we are interested in combining multiple eye behavior to improve a user intent detection method and investigating how the method helps solve Midas-touch.

1.5 Methodology

To answer the two research questions, we developed models that derived dwell time based on the relation between eye behaviors and human decision-making processes (Chapter 3) and detected user intent from eye behaviors using machine learning (Chapter 4). These models are based on the empirical data obtained through data collection experiments.

We addressed RQ1 by developing a model that derives dwell time. Previous research has determined dwell time to optimize the speed and accuracy of dwell selection. While dwell time plays an important role in user intent detection for

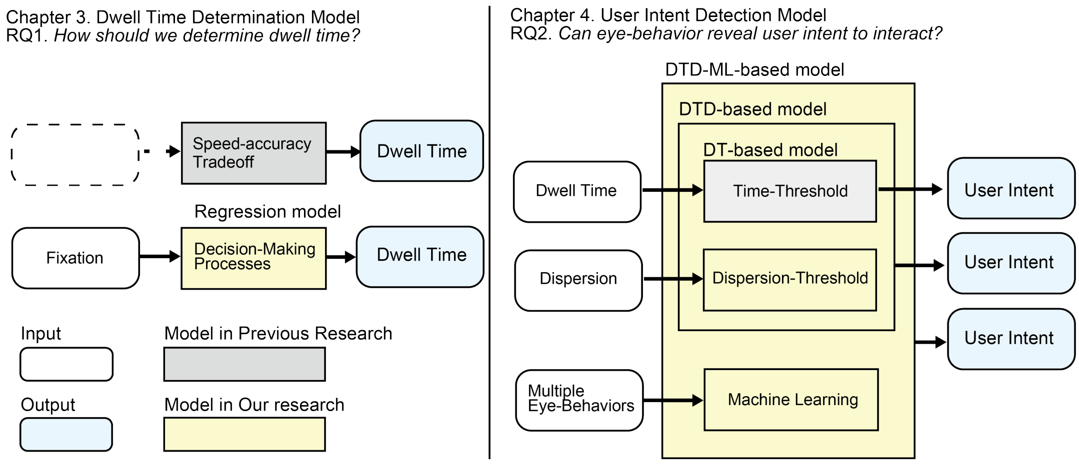

FIGURE 1.1: Overview of our work.

dwell selection, the method for determining dwell time does not involve the human decision-making process. We used the model human processor (MHP) shown by Card [CNM83]. The MHP is a well-known context in HCI that indicates a process until decision-making during a selection task against a visual stimulus. Because very few studies have explored a relation between dwell selection and MHP, we first explore the relation through user studies. Then, we developed the model using the relation (Figure 1.1, left).

We then addressed RQ2 by developing a user intent detection model. While intent detection for dwell selection has mainly relied on the time-based threshold (i.e., dwell time), we designed two intent detection methods (Figure 1.1, right). In this thesis, we refer to dwell time-based dwell selection as DT selection. One of our methods uses gaze coordinate dispersion in addition to dwell time; a dispersion of gaze coordinates during a dwell time is smaller than the dispersion threshold. By incorporating the dispersion threshold with dwell time, we aim to detect more careful user-looking action to prevent Midas-touch caused by mis-detection that the gaze coordinates are inside an object. We refer to this selection using dwell time and gaze dispersion as a dwell time-dispersion (DTD) selection. Furthermore, we used machine learning (ML)-based intent detection with multiple eye behavior for dwell selection in addition to DTD selection. We refer to this selection as DTD-ML selection. Lastly, we evaluated these two dwell selections to demonstrate that eye behavior can reveal user intent to select an object.

1.6 Contributions

The contributions of this thesis can be summarized as follows:

- **Model deriving dwell time based on the relation between fixation and the model human processor**. We demonstrate the relation between the natural human behavior of fixation, which indicates human attention and intention, and the human decision-making process, which is described by MHP. We develop our model based on the relation. Lastly, we show how the dwell time for five selection situations can be determined.

- **Model detecting user intent to select an object through multiple eye behavior and ML**. This model incorporates multiple eye behaviors to enhance the accuracy of intent detection in dwell selection. We demonstrate that this model can reduce the occurrence of Midas-touch, which is a long-standing issue in gaze-based interactions.

1.7 Thesis Structure

The remainder of this thesis is structured as follows:

Chapter 2 explains human natural eye behaviors used for gaze-based interaction. We then provide an overview of gaze-based interaction in various environments. Furthermore, we cover how researchers have determined dwell times and attempted to solve Midas-touch, which is the main challenge in dwell selection.

Chapter 3 introduces a study on determining dwell selection through a model that incorporates eye behavior and human decision-making processes. Based on the data obtained from experiments, we analyze human eye behavior during target selection tasks. The outcome is a detailed description of the relationship between the human natural eye behavior of fixation and the model human processor. The work presented in this chapter was originally published in the Proceedings of the ACM on Human-Computer Interaction (PACM HCI) [IYS23a].

Chapter 4 introduces a dwell selection using an ML model for user intent detection using multiple eye behaviors. Based on an experiment, we obtained data sets of eye behaviors during dwell selection with ground-truth labels of user intent, and developed an ML model using these datasets. Based on a comparison of the baseline dwell selection, we demonstrate the performance of our dwell selection with the ML model. The work presented in this chapter was originally published in Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (PACM IMWUT) [IYS22].

Lastly, we conclude this thesis by summarizing the study and describing the uses of our findings for future gaze-based interaction in **Chapter 5**.

Chapter 2

RELATED WORK

This thesis studies user intent detection to interact with a computer based on natural human eye behaviors. To contextualize our work, we first present types of eye behaviors used for gaze-based interaction. We then review gaze-based interaction research by first examining how users can interact with computers based on eye behaviors with three types of interaction methods. Because the focus of this thesis is on dwell selection, we also describe existing issues of dwell selection and how they have been addressed.

2.1 Eye Behaviors for Gaze-Based Interaction

Humans perceive visual stimuli through their eyes. They can see an object sharply when the light hits the fovea, which is a small region in the retina. Light enters through the cornea and the pupil, passes through the lens, and hits the retina. Eye movements help see an object sharply and ensure that the light directly hits on the small region of the fovea. The pupil regulates the amount by changing its size by contracting or relaxing the iris. These eye movements are aimed at controlling the amount of light, and focusing the lights on the fovea may be attributed to the movement of the extraocular muscles.

Although we cannot detect muscle movements through eye trackers, we can detect "gaze," which is a direction where users look. In a 2D display, gaze often represents the coordinations on the x-axis and y-axis; in a 3D environment, such as for an HMD, the gaze often represents directions to the x-, y-, and z-axes. The gaze itself does not have a powerful meaning as a cue of user attention and intent. Most eye behaviors used for gaze-based interaction are calculated using the gaze and timestamp when the gaze is sampled. Further, there is no strong definition of the calculation; it depends on what researchers want to know from eye behaviors. In the following sections, we review such eye behaviors (Figure 2.1) to deepen our understanding of how gaze-based interaction is developed.

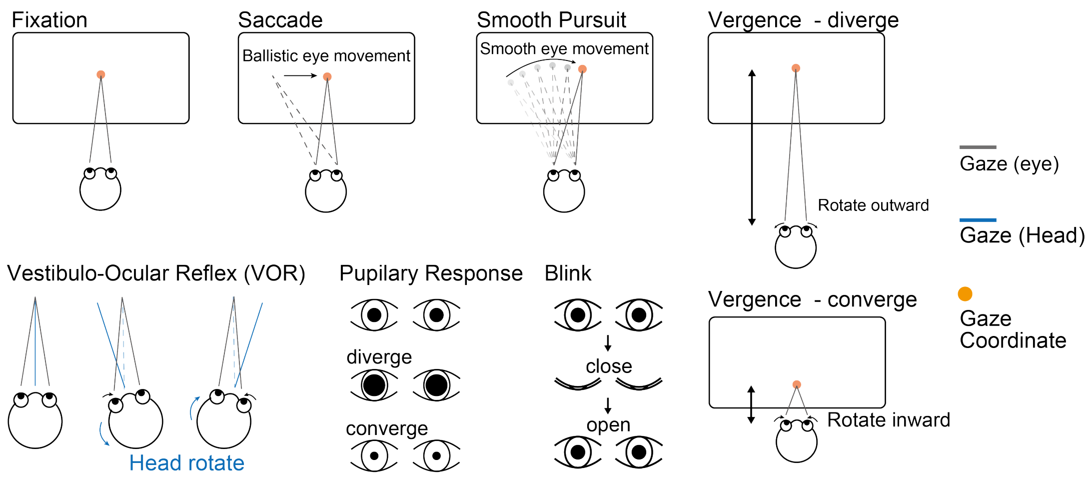

FIGURE 2.1: Eye behavior used for gaze-based interaction.

Saccade is a rapid and conjugate (both eyes do the same thing) eye movement that humans make when re-orienting the fovea to a new spatial location. Saccades are often described as ballistic eye movements, meaning that their direction is not altered once they start moving the eyes. Moreover, based on recent eye-tracking research, humans are generally assumed to be blind during saccades; many researchers focus more on what is being looked at and how long it was looked at for, and hence, the saccade is not widely adopted as a cue for user attention and intention. However, their duration, amplitude, direction, and peak velocity have been used to indicate how users move their eyes or how their attention shifts.

Fixation is a stable eye movement wherein the eyes almost stop allowing the light hit the fovea. Humans can take in detailed information on what is being looked at by fixation. A fixation is typically done before or after saccades and is often used as a cue of human attention to analyze human behavior. The important metrics are location (can be derived from gaze) and duration (can be derived from timestamp) of fixation, meaning how long a human attentively looks at a point. The duration of fixation is typically between 200–600 ms, but it can be much shorter or longer.

The eyes move slightly during fixation; the eye movements are referred to as microsaccade, which is a small saccade, a drift, which is a slow change in location, and a tremor, which is a muscle contraction and relaxation. Such small eye movements occur in a fixation. The microsaccade is the largest movement, with less than 0.4 degrees while the tremor is the smallest eye movement, with approximately 0.004 degrees. While the microsaccades can be detected using a high-performance eye tracker (e.g., the Tobii Pro Spectrum by Tobii and the EyeLink 1000 Plus by SR Research Ltd.), which has a fast sampling rate and high level of precision, it is difficult to detect the tremors using such eye trackers.

In the HCI field, saccades and fixations are often detected using the algorithm proposed by Salvucci and Goldberg [SG00]. They proposed three algorithms based on velocity, dispersion, and area. The velocity-based algorithm is referred to as Velocity-Threshold Identification (I-VT) algorithm. This algorithm uses the two velocity thresholds: low velocities (e.g., <100 deg/s) for fixation detection and high velocities (e.g., >300 deg/s) for saccade detection. The dispersion-based algorithm is referred to as Dispersion-Threshold Identification (I-DT) algorithm, and it detects only fixation with a dispersion threshold for user gaze coordinates. The I-DT algorithm requires two thresholds for duration and dispersion. The dispersion threshold is 0.5°–1.0°. The duration threshold varies between 100–200 ms [Wid84], depending on the tasks. The centroid of gaze coordinates is regarded as the fixated point. The area-based algorithm is referred to as Area-of-Interest Identification (I-AOI) algorithm. In contrast, the I-AOI algorithm only identifies fixations within the specified target area. The fixations are detected with a duration threshold similar to the I-DT algorithm. Other eye movements outside the specified target are regarded as saccades. Those thresholds are determined based on interchanging for each research.

Smooth Pursuit is a smooth eye movement for following a moving object that humans originally fixated on. This allows humans to maintain the light hitting on the fovea, which causes the light to move. Similar to saccade and fixation, smooth pursuit is also naturally done by humans. If humans follow a moving object with a saccade, the light hits on the fovea is reduced because humans are assumed to be blind during the saccade. A smooth pursuit is often detected by calculating the correlation coefficient of the gaze coordinates and the moving object [VBG13].

Vestibulo-ocular reflex (VOR) is an eye movement that follows the head movement. VOR ensures that the light keeps hitting the object on which humans originally fixated on the fovea even if the head moves, by that the eyes rotate in the opposite direction to the head rotation.

Vergence is the movement of the eyes in the opposite direction when shifting the focus between near and far objects. During vergence, the eyes rotate inward and outward to focus on near and far objects, respectively. This allows humans to take light on the fovea from a distant point. In HCI research, a vergence is often detected by calculating the difference in pupil positions of both eyes or the difference in gaze directions on the x-axis of both eyes or inter-pupillary distance.

Pupillary Response can often be shown when the pupil regulates the amount of light hitting on the fovea. Therefore, this generally depends on the ambient light conditions. If the light is strong (e.g., under the sun), the pupil contracts to decrease the amount, and in dark conditions, the pupil relaxes to take in a large amount of the light [HP60]. Moreover, it is known that this response is affected by human interest and emotion [HP60].

Blinking technically is not a movement of the eye itself but a semi-autonomic rapid eyelid closing. However, blinking has an important role in hitting the light on the fovea by clearing away particles from the eyes and lubricating the eyeballs to maintain the eyes to ensure a smooth surface. The maximum duration of a single blink of closure duration is 500 ms [CEU03, SGBG08].

2.2 Gaze-Based Interaction

Gaze-based interaction uses human eye behaviors of looking at a point (fixation) and moving the eyes (saccade) by interpreting that these eye behaviors have a role of indicating attention and shifting the attention towards another point, respectively. Therefore, researchers have developed various gaze-based interactions. The first research to develop a selection method was conducted by Ware and Mikaelian [WM87]. They explored how left-clicking in mouse-based interaction can be imitated using eye behavior. Furthermore, researchers have extended gaze-based interaction to involve the command triggering method and assistive role for other modalities. We categorize gaze-based interaction into three types: implicit, explicit, and multimodal gaze interactions.

2.2.1 Implicit Gaze Interaction

Our eyes are implicitly used to indicate attention and intent and to guide action for ourselves. As we reach for and interact with objects in the physical world through our hands, our gaze first moves to the object, and we move our hands toward the object. Similarly, to interact with GUI objects, our gaze first moves toward the objects, and we then move the cursor toward the object. Indicating our attention to others is attractive for incorporating eye behaviors into interaction, and the implicit gaze interaction is based on this factor. Numerous studies have used the user gaze as a cursor and other modalities for triggering interaction; we explain such interaction as multimodal gaze interaction later.

As an implicit gaze interaction, dwell selection, which utilizes the natural eye behavior of looking as both cursor and modality for triggering selection, is the most researched interaction method. The natural eye behaviors are used for an attentive user interface, automatically adapting the user interface for user attention [Ver03]. For example, an interface that shows translation according to a user reading eye behavior [HMAR00] and suggests self-confidence to the users to help in decision-making [IMKD20] is studied. Several researchers have used natural eye behaviors of fixations, saccades, and pupillary changes to detect attention (e.g., [AL13, XSB16]), cognitive states (e.g., [HB05]), and decision intent (e.g., [LH01, JSSV15]). Such detection has been used for gaze-based interaction and designing a web page or visualizing user intent.

2.2.2 Explicit Gaze Interaction

There are various types of explicit gaze interactions, including voluntary *saccade*, *vergence*, *pursuit*, and *eye-gesture*. While these eye behaviors are natural eye behaviors, explicit gaze interaction utilizes user voluntary eye behaviors. These voluntary eye behaviors are used as an indicator of user intent. Because voluntary eye behaviors used for explicit gaze interaction are designed to distinguish from natural human eye behaviors, Midas-touch rarely occurs. We described the explicit gaze interaction with three interaction methods by integrating the confirmation button, moving object and smooth pursuit, and eye-gestures.

Interaction with Confirmation Button

The confirmation button, an additional arranged button, which allows a computer to detect user intent to select the button, has been adopted to prevent Midas-touch. Ware and Collins [WM87] conducted the first research that adopts the confirmation button, which pops up next to the target after looking at the target. The users can select the target by first looking at it, moving their eyes to its confirmation button, and then dwelling on it; that is, this interaction requires a saccade in addition to dwell selection. Significant research is conducted toward improving interaction with the confirmation button [MGFY18, LPW15, SLW19, FF18, PLW13, WM87, SRT11, CSO22]. As an additional saccade is incorporated, the dwell time for dwell selection on the confirmation button can be smaller than for an original dwell selection. However, placing the confirmation button next to the GUI object causes unwanted selections considering that users may accidentally look at the confirmation button. As the interaction method extends the confirmation button, the ActiGaze [LPW15] principle is applied. Following this principle, potential targets are colored, and the colored confirmation buttons corresponding to the color of targets are arranged at the periphery of users (e.g., side of the display). Users can select the desired target by dwelling on the confirmation button, whose colors correspond to the target to prevent unwanted selection; unless users look at the confirmation button, no selection is triggered. Moreover, in the ActiGaze principle, the multiple confirmation buttons potentially solve the issue of selecting a small and/or closed target, which is another issue of gaze-based interaction caused by the low eye tracker performance. This is because the size and arrangement of the confirmation buttons can be freely determined, and the confirmation buttons are generally larger than the target and are arranged with some margins between buttons.

While the above confirmation button requires a horizontal and vertical gaze movement (i.e., saccade) for selection, selection with a vergence, which is a gaze movement for the depth direction is also explored. Users can select the target by first looking at it and then refocusing and dwelling on the confirmation button, which is arranged either behind or in front of the display [KB16, KOH+13].

Interaction with Smooth Pursuit

In gaze-based interaction, smooth pursuits are induced by the motion of an object, which is the target itself [VBG13] or additionally arranged moving object(s) [VCKM18, SDRD17, DKA18, EVBG15, ŠIK+16, DHI17, ASLL20, SCN+23]. Users can trigger an interaction by constantly looking at a moving object. The computer measures a correlation between the target and eye movement to detect user intent in order to interact with the target and identify the target that the users follow. If a command is assigned to a moving object (a command name is labeled), detecting smooth pursuit for such objects triggers a command [DHI17]. This is because the interaction is not triggered unless users keep looking at moving objects, and thus, Midas-touches can be solved. Moreover, the issue of selecting a small and/or closed target caused by a low eye-tracking performance is solved as this interaction only uses the correlation, and the absolute gaze coordinates are not necessary. These properties are useful for gaze-based interaction for a small device such as a smartwatch [EVBG15].

Interaction with Eye-Gesture

Eye-gesture is designed to trigger commands, such as “copy” and “paste.” The gestures are determined beforehand and users can trigger a command by moving their eyes to form such gestures. The most simple gesture is the one that uses a single stroke (right-to-left or left-to-right) of eye movement [MHL13b, MHLG09, MHL13a, MLGH10, RH18]. In terms of robustness against unwanted interaction, a gesture comprising two or more strokes of eye movements [DS07, IHI+10, WRSD08, IYS20] is better. However, it is difficult to move the eyes on a larger number of strokes or at a larger distance owing to its complexity. Therefore, visual guidance (e.g., a menu) has been adopted to help users easily trigger a command; for example, an additionally displayed window [WRSD08], a semi-transparent region [ULH10], or a physical object [JHF17]. Moreover, a combination of dwell selection and eye-gesture is proposed. For example, by using a pie menu, the menu is displayed after fixation, and the command is activated when the gaze crosses the edges of the menu [HU08, ULH10, ASP+21a]. Similarly, several eye-gestures for the marking menu [Kur93] are researched [KHAL22a].

Moreover, interaction using blink and wink is studied. In contrast to the above eye-gestures, which are for triggering a command, voluntary blinking and winking are used as eye-gestures to trigger a selection. A computer detects user intent to interact with a computer through the eye-gesture of closing and then opening the eyes; such a gesture is used as a cue for detecting the user intent. For example, users can trigger a selection by a voluntary blink [GBL+03, KS19, MB10]. Because blinking (closing and opening both eyes) is a natural human eye behavior, accurate voluntary blinking detection has been researched, similar to dwell selection. Researchers have explored a duration threshold that distinguishes natural and voluntary blinking for interaction. For example, because the closure duration of a single blink requires at most 500 ms [CEU03, SGBG08], a duration longer than 500 ms is used for detecting a voluntary user blinking [GBL+03]. However, a duration over 500 ms of eye closure can be regarded to as the natural human eye behavior of microsleep; therefore, it can be difficult to prevent mis-detection with only a duration threshold, similar to dwell time. Recent research utilized winking, the gesture of closing and opening one eye [RGCSG21]. Compared to blinking, winking rarely occurs in natural eye behavior. Thus, to detect winking, a duration threshold of 250 ms is used [RGCSG21], which is smaller than that used for detecting blinking. Because humans can voluntarily open and close each eye and keep either eye closed while gazing with the other [JW15], moving one eye while closing another eye allows for mimicking a mouse-based interaction of “drag and drop” [RGCSG21]. Users can hold a target by looking at it and then closing one eye; they can then drop it by opening the closed eye after moving another toward the desired position.

2.2.3 Multimodal Gaze Interaction

In contrast to implicit and explicit gaze interactions, which utilize eye behaviors as the primary modality, multimodal gaze interaction employs eye behaviors as an assistive modality alongside other modalities. In multimodal gaze interaction, eye behaviors are used as a cursor and other modalities are used as a cue to indicate the user intent. Similar to the explicit gaze interaction, the Midas-touch rarely occurs.

Incorporating eye behaviors into hand-based interaction, which uses a mouse, touchpanel, and hand gestures, is researched. The first work to incorporate natural eye behaviors with hand-based interaction is the MAGIC pointing proposed by Zhai et al. [ZMI99]. They argued that "it is unnatural to overload a perceptual channel such as vision with a motor control task." As a result, they proposed a pointing method that replaces moving a cursor with the mouse by looking at the desired point on display. The experiment indicated that MAGIC pointing could reduce physical effort compared to mouse-based interaction.

The original work of MAGIC pointing is aimed at mouse-based interaction, which has been extended in various situations and hand-based interactions. By adopting the MAGIC pointing for gaze-based interaction, manipulating smartphone and tablet devices with touch-based or pen-based interaction with eye behavior is shown [PACG14, SD12b, KAH+16, PAC+15, NSA+23]. For mouse- and touch-based interactions, the primary role of eye behavior in MAGIC pointing is to make the interaction faster. Moreover, eye behavior allows users to interact with distant objects, such as a large tablet, public displays, and a virtual reality environment [SD12a, SD12b, SD13, TABG15, PACG14, PMMG17, CXH15]. For gaze-based interaction in HMD-based interaction, the use of VOR has been studied [SG19b, SG19a, PLLB17]. These multimodal gaze interactions have higher interaction performance than implicit and explicit gaze interactions [CXH15, SG19b].

Similarly, an implicit use of eye behavior as a cue of user attention to support voice-based interaction is used [MLH20, KNBV22]. Current voice assistants, such as Alexa by Amazon and Siri by Apple, do not use contextual information regarding user attention while users speak commands. Applying the implicit use of eye behavior as contextual information for voice assistants has been studied [MLH20, KNBV22]. Because both gaze and voice are ambiguous in user attention and spoken command using each as a stand-alone, mis-detections tended to occur. A combination of both could resolve the ambiguities and enable faster interaction [ZIGM04].

As highlighted by the aforementioned studies, complementing the gaze with some other modalities can extend current interaction to more useful interaction that cannot be established using current primary modalities such as only hands.

2.3 Dwell Selection Challenges

In this thesis, we address two research questions related to eye behaviors and user intent in the context of dwell selection (RQ1 and RQ2). We introduce research that addresses two questions.

2.3.1 Dwell Time for Dwell Selection

A majority of the studies on exploring dwell time focused on optimizing dwell time to achieve fast and accurate dwell selection. Researchers on dwell typing (i.e., dwell selection on a key) used dwell times between 180–600 ms from two perspectives: user preference and robustness against the Midas-touch [MAv09, RO12, PS17, NDA+17, vM04]. To select a key that is most likely to be selected, a small dwell time is ideal since it enables a faster selection, while using a large dwell time prevents Midas-touch for a key that is unlikely to be selected. Dwell times dynamically decrease/increase along with the previously typed keys and the probability of the next typed key.

Although dwell time plays a significant role in intent detection for dwell selection, there is a lack of human decision-making processes in the existing determination approaches for dwell time. However, although the concept of incorporating the human decision-making process aims for a faster selection, the difference from the current dwell time determination method is that the dwell time should not be considerably small. For example, we previously found that using a small dwell time (100 ms for selecting a simple colored object) decreases usability from the questionnaire in the experiment; participants answered *I felt that the target was acquired before I looked at the target* [IAST18]. Another previous research reported that a participant answered *When the dwell time was too short, the selection was completed before I could recognize the panel.* [CSO22]. This suggests that even if Midas-touch is entirely solved, a smaller dwell time or 0 s is not always an optimal solution, which contradicts the previous research pursuing a smaller dwell time for preventing Midas-touch.

Dwell time should also be determined for the dwell selection used to compare the performance with other interaction methods, such as dwell selection vs. eye-gesture or dwell selection vs. multimodal gaze interaction (e.g., [SCN+23, CSO22]). Researchers have adopted their own dwell time that is determined by referring to previous research or conducting the preliminary study. However, a detailed description has been skipped in their research. Because eye behaviors vary according to tasks, conditions, and situations, the dwell time should also be carefully determined.

2.3.2 Midas-touch and Solution

Gaze-based interaction has faced Midas-touch, an unwanted selection. The origin of the phrase "Midas-touch" is in Greek mythology about King Midas for his ability to turn everything he touched into gold. Midas-touch occurs owing to the difficulty in accurate user intent detection. Solving Midas-touch has always been the focal topic of gaze-based interaction research.

Many researchers have attempted to detect the user's intent to prevent Midas-touch. One approach aims at adjusting the dwell time by making it larger or smaller according to the situation or users. The easiest solution is to use a longer dwell time; however, this solution decreases usability. Moreover, 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. Therefore, researchers 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 [MAv09, RO12, HJH+03, MAR04, MMAR06, MWWM17, PSD12, PS17]. Another approach is to adjust the dwell time according to the target [NDA+17] or the eye movement before landing on the target [IAST18]. However, even though the task's cognitive load strongly affects the occurrence of Midas-touches [ZXZZ11], these studies are aimed at selecting colored targets or simple images. A few studies have developed an ML-based intent detection system using eye behaviors of fixations, saccades, and pupillary response during selection tasks [BVH12]. However, because their ML-based system requires eye behaviors after the selection is triggered, the system cannot work in a real-time interaction.

2.4 Position of This Thesis

Each gaze-based interaction has certain advantages. Implicit interaction can benefit from using natural eye behaviors for gaze-based interaction, which is accessible for various users and fast interaction. Explicit interaction allows users to trigger various interactions, not limited to selection. Triggering various interactions is important to enable gaze-based interaction toward everyday interaction. Multimodal gaze interaction utilizes eye behaviors to extend current interaction methods, thereby improving the usability of those interactions. Studying all interactions is indispensable for extending the current interactions for more users and situations; there is no unique and best interaction method, and they should be improved in each and reciprocally. Because all gaze-based interactions are based on an implicit use of eye behavior that indicates a cue of user attention and intent, the precise detection of user attention and intent is important for improvement across whole gaze-based interaction. In this thesis, we aim to extend gaze-based interaction by developing the user intent detection method through natural human eye behavior for implicit gaze interaction of dwell selection.

To improve the dwell selection, we find two aspects through previous research. First, the determination of dwell time, an indispensable parameter of dwell selection, is based on the speed and accuracy of selection. While dwell time plays an important role in user intent detection for dwell selection, the determination method of dwell time lacks the incorporation of the human decision-making process. Second, solving Midas-touch has primarily relied on dwell time and eye movement, although there are other eye behaviors indicating user intent. In this thesis, we investigate eye behaviors during selection tasks and leverage multiple eye behaviors to tackle fundamental dwell selection challenges.

Chapter 3

DETERMINATION OF DWELL TIME THROUGH FIXATION AND MHP

This chapter explores the determination of dwell time that incorporates natural human eye behavior and the human decision-making process. We focus on fixation, which indicates user attention, and the MHP [CNM83], which shows human perception and cognition processes in decision-making.

We develop a model that derives dwell time and allows dwell selection after a user completes the decision-making process based on their behavior. We first propose three hypotheses regarding the relations between the fixation information and the decision-making process. Based on the experimental findings, we justify those hypotheses and develop our model to derive the dwell time using the number of fixations that a user performs for a target ($\mathrm{N}_{\mathrm{fixation}}$) and the duration of fixations that a user performs for a target ($\mathrm{D}_{\mathrm{fixation}}$). During the experiment, we measured $\mathrm{N}_{\mathrm{fixation}}$ and $\mathrm{D}_{\mathrm{fixation}}$ for an instructed target. Because the decision-making process varies for different tasks, we conducted five selection tasks with different difficulties, and we then evaluated our hypotheses and developed our model.

In this chapter, we first describe the human decision-making process that is interpreted by the MHP [CNM83], propose hypotheses on the relation between fixation and MHP, validate the hypotheses through experiments, develop our model, and demonstrate the applications of our model.

The contributions of this work are summarized as follows.

- We proposed three hypotheses about fixation during selection and validated them through an experiment involving five tasks with different difficulties.
- We developed a model that derives dwell time and allows dwell selection after a user completes the decision-making process by incorporating the natural human eye behavior of fixation and the human decision-making process.

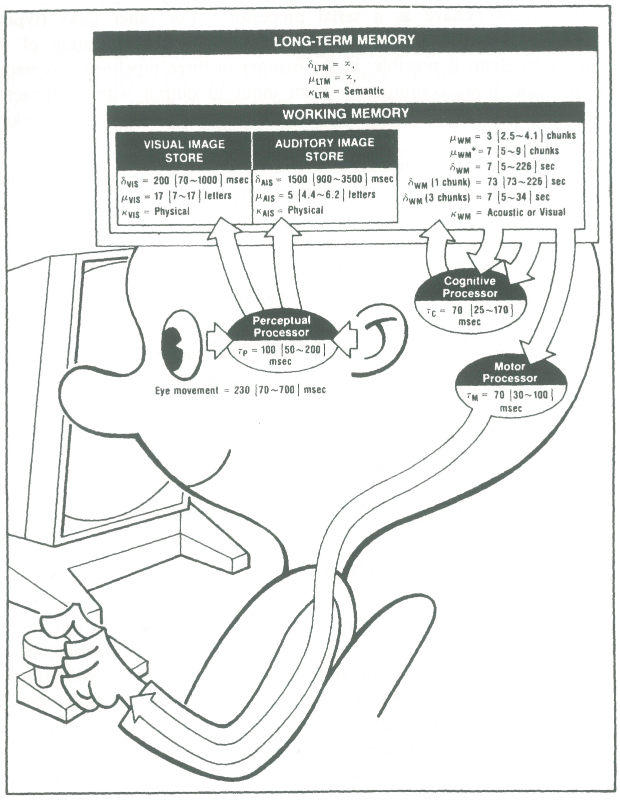

FIGURE 3.1: Overview of MHP. Image from Card et al. [CNM83].

- We showed how our model derives dynamically changing dwell times based on user behavior, especially $N_{fixation}$.

3.1 Human Decision-Making Process via Model Human Processor

The MHP demonstrates human perceptual behavior in response to the visual (and auditory) stimulus by dividing the information-processing system into three subsystems: perception, cognition, and motor systems (Figure 3.1). The perception subsystem completes *perceiving* a visual stimulus and encodes it into a visual code within τ_p=100 [50–200] ms. Each range indicates that the Fastman (e.g., an expert) takes the minimum time, and the Slowman (e.g., a novice) takes the maximum time. The cognition subsystem completes *recognizing* the visual code, *classifying* the recognized code into a meaning, *matching* the meaning and instruction loaded on the

TABLE 3.1: Tasks and their instructions described in [CNM83].

Task	Instruction (Push a button if...)	Example
Simple reaction	an **object is displayed**	-
Physical match	the **shape** of the object is correct	‘a’ v.s. ‘a’
Name match	the **name** of the object is correct	‘a’ v.s. ‘A’
Class match	the **content** of the object is correct	letter vs. letter

TABLE 3.2: Tasks and their required processes described in [CNM83].

Task	Required process					
Simple reaction	*perceive*				*request*	*act*
Physical match	*perceive*			*match*	*request*	*act*
Name match	*perceive*	*recognize*		*match*	*request*	*act*
Class match	*perceive*	*recognize*	*classify*	*match*	*request*	*act*

working memory beforehand, and *requesting* to *act* process to the motor subsystem. Therefore, the *request* process can be regarded as the process of decision-making for tasks described in [CNM83]. The time taken for one cognitive process (τ_c) is 70 [25–170] ms. The motor subsystem completes *acting* (i.e., pushing a button for tasks described in [CNM83]) along with the request from the cognition subsystem within τ_m=70 [30–100] ms.

3.1.1 Selection Tasks in MHP

Card et al. [CNM83] described the required decision-making process for completing a task, which differs between tasks, and gave examples of four selection tasks (Table 3.1) and their required processes (Table 3.2). The tasks are consistent in the sense that participants push the button located under their hand in response to the visual stimulus (an object) shown in the display. Here, the differences are in the task instructions. The simplest task in [CNM83] is the simple reaction task, where the instruction is to push a button when an object is displayed. The required processes are *perceive, request*, and *act* because participants only make a decision when they perceive a visual stimulus; thus, no further cognitive subsystem is required. The second simplest task is the physical match task, where the instruction is to push a button if the shape of the object matches the instruction. The required processes are *perceive, match, request*, and *act*. For example, if the instruction is ‘a’ and the stimulus is 'a,' then the participant should press a button; if the instruction is 'a' and the stimulus is 'A,' then the participant should not press a button; participants require a *match* process to match whether both stimuli are same (i.e., the physical shapes are same in this case) or not in addition to the required task of the simple reaction task. The third simplest task is the name match task, where the instruction is to push a button if the name of the object matches the instruction. For example, because the names 'a' and 'A' are both 'a,' if the instruction is 'a' and the stimulus is 'A,' participants push a button. The required processes are *perceive, recognize, match, request*, and *act*. In addition to the required processes of the physical match task, humans are required the *recognize* to recognize the objects (i.e., the name of stimuli in this case). The most difficult task is the class match task. Here, the instruction is to push a button if the class of the object matches the instruction. For example, as 'a' and 'b' are both letters, if the instruction is 'a' and the stimulus is 'b,' participants push a button; conversely, if the instruction is 'a' and the stimulus is '3,' the participants should not push a button. The required processes are *perceive, recognize, classify, match, request*, and *act*. In addition to the processes required for the name match task, the *classify* process is required for the classification of the objects (i.e., the class (image, letter, or number, as well as the image of dog, cat, bird) of the stimuli.

3.1.2 Time Required for Competing Tasks

Because the MHP describes the required processes and requires considerable time to complete one process, we can estimate the duration from the beginning of perceiving a stimulus to the end of pushing a button. The time for completing the simple reaction task, whose required processes are *perceive, request*, and *act*, is:

$$240\ [105{-}470] = \tau_{\mathrm{p}} + \tau_{\mathrm{c}} + \tau_{\mathrm{m}} = 100 + 70 + 70.$$

The time for completing the physical match task, whose required processes are *perceive, match, request*, and *act*, is:

$$310\ [130{-}640] = \tau_{\mathrm{p}} + 2\tau_{\mathrm{c}} + \tau_{\mathrm{m}} = 100 + 140 + 70.$$

The time for completing the name match task, whose required processes are *perceive, recognize, match, request*, and *act*, is:

$$380\ [155{-}810 = \tau_{\mathrm{p}} + 3\tau_{\mathrm{c}} + \tau_{\mathrm{m}} = 100 + 210 + 70.$$

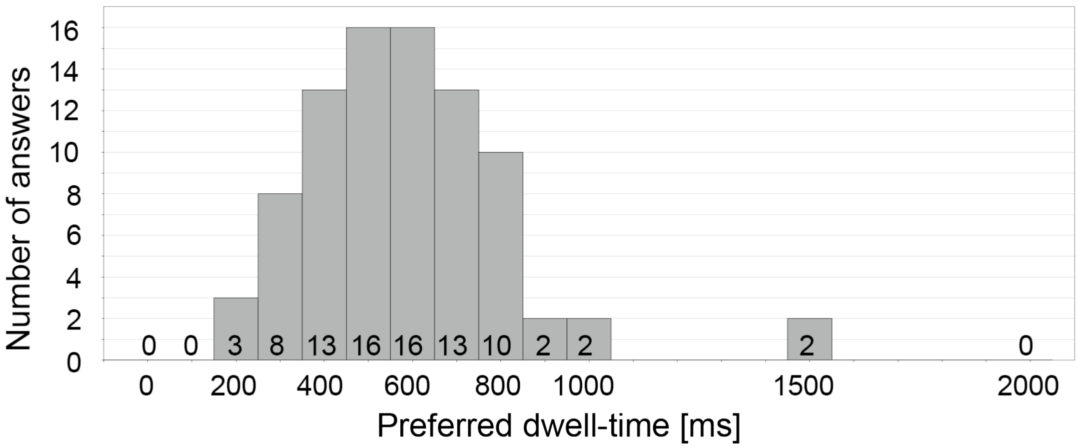

FIGURE 3.2: Participants' preferred dwell time for image selection task in our previous work [IYS21].

The time for completing the class match task, whose required processes are *perceive, recognize, classify, match, request* and *act*, is:

$$450\ [180{-}980] = \tau_{\mathrm{p}} + 4\tau_{\mathrm{c}} + \tau_{\mathrm{m}} = 100 + 280 + 70.$$

All tasks require *perceive, request*, and *act*. Therefore, the main difference among tasks lies in the required processes on the cognition subsystem of *recognize, classify*, and *match*.

3.1.3 Relation Between Dwell Time and MHP

We previously showed the relation between dwell time and MHP [IYS21]. We asked 16 participants to complete an image selection task that imitates a class match task in [CNM83] with two selection methods: gaze-button and dwell selections. In the gaze-button selection, users can select a target by looking at and pushing an enter key on a keyboard placed at the participant's hand. Through the experiment, we observe two times: the button press time and user-preferred dwell time. The button press time is measured from when the participant's gaze enters a target to when the participant pushes a button to select. Because we imitated the task as a class match task in [CNM83], the button press time ideally equals the time required to complete the task. The preferred dwell time is obtained by asking participants their preferences for each dwell time after they try all dwell times of 100, 200, 300, ..., 1000, 1500, and 2000 ms. For example, we asked, "Do you prefer

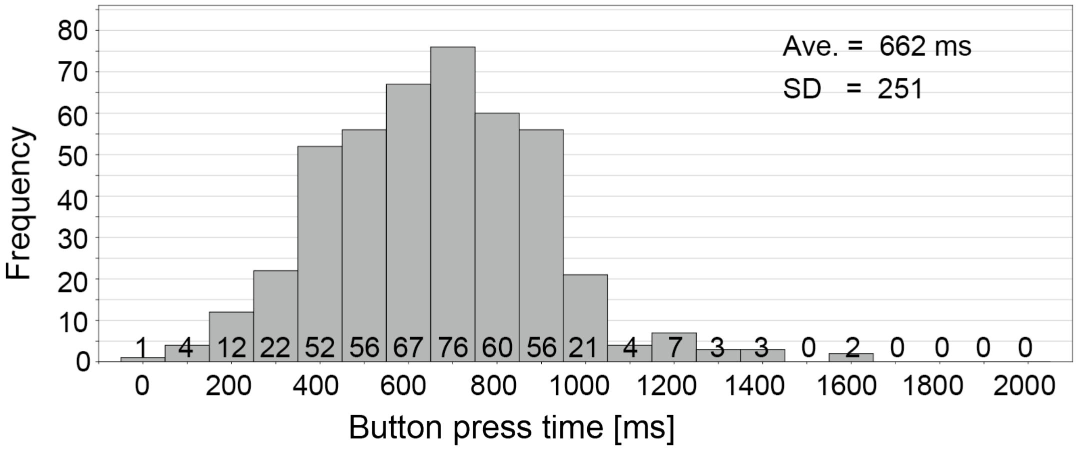

FIGURE 3.3: Button press time for image selection task in our previous work [IYS21].

xx ms as dwell time?" Through analysis, we found the following. First, all participants preferred 500 ms and 600 ms as a dwell time for an image selection task as shown in Figure 3.2. Second, the button press time averaged 662 ms (SD=251) as shown in Figure 3.3. Third, the number of fixations that participants perform for a target during the selection task averaged 2.30 (SD=0.82); 11.4%, 56.4%, 24.8%, 5.8%, and 1.5% of button selections are completed with one, two, three, four, and five fixations, respectively.

In the MHP, the *requesting* process on the cognition subsystem can be regarded as the process of decision-making, and the *act* process on the motor subsystem is not involved in the decision-making processes. Therefore, we consider that τ_m is subtracted from button press time as the time required for the decision-making; the decision-making requires 592 ms (=662−70 ms). The difference between participants' preferred dwell time and the time required for decision-making seemed to be caused by the required duration range for each MHP process.

The required processes for completing the image selection task that we imitate the class match task in [CNM83] are *perceive*, and $\mathrm{N}_{\mathrm{fixation}}$ times of *recognize, classify*, and *match*. The difference in required processes among tasks shown in [CNM83] is the number of required processes for the cognition subsystem, as shown in Table 3.2. Considering this and the experimentally observed button press time and the number of fixations, we showed the first model determining dwell time from $\mathrm{N}_{\mathrm{fixation}}$ as:

$$\tau_p + (3\mathrm{N}_{\mathrm{fixation}} + 1)\tau_c. \tag{3.1}$$

This model can be interpreted such that the dwell time should include the time required for *perceive*, N_{fixation} times of *recognize, classify*, and *match*, and *request*. Note that because pushing a button is not necessary for completing the task with dwell selection, the time required for *act* process is not counted in the model. With this model, we can determine the dwell time for an image selection task with predicted N_{fixation} required for completing the task, as shown below.

$$\begin{aligned}
380\,\text{ms}\ [150{-}880] \quad &= 100 + (3 \times 1 + 1) \times 70\ (N_{\text{fixation}} = 1),\\
590\,\text{ms}\ [225{-}1,390] &= 100 + (3 \times 2 + 1) \times 70\ (N_{\text{fixation}} = 2),\ \text{and}\\
800\,\text{ms}\ [300{-}1,900] &= 100 + (3 \times 3 + 1) \times 70\ (N_{\text{fixation}} = 3).
\end{aligned}$$

Because half or more participants preferred dwell times of 300–800 ms, the dwell time determined through this model that ranged into 380–800 ms seemed to be fitted. In other words, by using this model, it may be possible to determine user-preferred dwell time according to the decision-making processes.

This previous work is the first work to explore a relation between fixation and the human decision-making process using the MHP. However, N_{fixation} prediction required for completing a task is challenging, and the applicable task is only for image selection (i.e., the class match task in MHP). In this thesis, we improved our previous model to be applicable for five selection tasks on five different targets: a simple colored object, letter, key, word, and image.

3.1.4 Models of Human Cognition and Behavior for Visual Search Tasks

In many studies in the HCI field, human cognition and behavior were modeled in a manner similar to [CNM83]. For instance, the adaptive control of thought–rational (ACT-R) model [AML97, AMD95] is a representative model of the human cognition process, including visual attention. The ACT-R model interprets that a human takes 186 ms to shift attention with or without eye movement. In a visual search task, three processes occur repeatedly: 1) responding "yes" (i.e., a looking candidate is a target), taking a "base" time of 208 ms, 2) shifting attention, taking a "shift" time of 186 ms, and 3) responding "no" (i.e., there is no target after searching all candidates), taking a "base" time and "neg" time of 133 ms (i.e., 208 + 133 = 241 ms)[1]. Another representative model is Fitts' law [Fit54, Mac91], which is aimed at pointing behavior. The time for pointing is expressed as $a \times \log_2(A/W + 1) + b$, where A is the distance between the position of a cursor and target, W is the target size, a is the time required for the motor process (e.g., moving a hand for

[1]These values depend on the difference in the types of targets and distractors (e.g., letters versus numbers) and the number of candidates present in a visual search task.

a mouse-based interaction), and b is the time required for the decision-making and triggering action. Moreover, numerous models have been proposed for GUIs (e.g., [CGG07, BOBH14, PL18]).

These models provide a precise representation of human cognition processes and behaviors, including the time required for each process. In this work, we adopt the MHP to explore a new dwell time determination method for the following reasons. We can interpret the human decision-making process through six processors based on three subsystems by using MHP. The required duration for all processors is reported in [CNM83]; τ_p for the process in the perception subsystem, τ_c for the processes in the cognition subsystem, and τ_m for the process in the motor subsystem. Card et al. [CNM83] describe the required processes for completing four tasks. This description is also useful for determining dwell time against various tasks, as Zhang et al. [ZXZZ11] reported that the dwell time should be determined for each task.

3.2 Hypotheses

We first propose the following hypotheses to understand the relationship between the natural human eye behavior of fixation and the human decision-making process. We mainly focus on the fixation information of $\mathrm{N}_{\mathrm{fixation}}$ and $\mathrm{D}_{\mathrm{fixation}}$.

H1. *N_{fixation} required for selecting a target increases along with the difficulty of our task.* We assume that users need to fixate on the target several times for completing more difficult tasks, which includes selecting a more complex target, before deciding to select it.

H2. *D_{fixation} of the fixation, when the target is selected, decreases as total N_{fixation} increases.* We assume that users can decide to select the target by fixating on it for a shorter period when they previously fixated on the target many times and recognized the target beforehand.

H3. *D_{fixation} for large N_{fixation} converges to the duration required for completing decision-making processes for a simple reaction task regardless of the task.* We assume that if a user has already recognized a target, they can make a decision to select the target with the duration regardless of the target type. In particular, the time required for decision-making converges to the duration required to complete decision-making processes for a simple reaction task, which is the easiest task in the MHP.

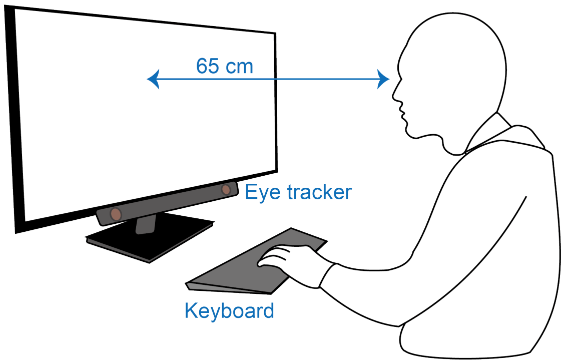

FIGURE 3.4: Experimental environment.

3.3 Experiment

We used five selection tasks with different difficulties to verify the hypotheses and determine the N_{fixation} required for a selection and D_{fixation}.

3.3.1 Participants and Apparatus

We recruited 20 university students (one female and 19 males, all Japanese) aged 20–26 ($M = 22.9$). They used GUI-based interfaces daily. Fifteen of them previously participated in an experiment using an eye tracker. Each received JPY 2,500 (∼USD 18).

We used the Tobii Pro Spectrum, which samples gaze data at 1200 Hz (0.833 ms/sample) with an accuracy of 0.6° and a precision of 0.06°. The eye tracker was attached to the bottom of a 24 inch (1920 × 1080 pixels) non-glare display. The participants' heads were positioned 65 cm away from the display. The participants used a wire-connected keyboard to control the task. The experimental environment is shown in Figure 3.4. The experiment was conducted in a room with fluorescent light at approximately 810 lux.

3.3.2 Selection Method

We used *gaze-button* selection, which is performed on the gaze coordinate when pushing the 'Enter' key of the keyboard. Selection is allowed when the gaze coordinate is inside a target; else, no selection is performed even if the participants push the key.

TABLE 3.3: Tasks and their difficulties. "Known candidates" means whether or not a participant knew which keys/icons/words/images were shown in candidates before a task began. We assigned "difficulties" in accordance with the row "Similar task in MHP," where each task requires a different number of required decision-making processes. The difference in "Known candidates" between the key and icon tasks results in a different difficulty even though the task in the MHP is the same.

Tasks	Target type	Known candidates	Similar task in MHP	Difficulty of task
Simple	colored object	beforehand	simple reaction	1 (minimum)
Key	key	beforehand	physical match	2
Icon	desktop icon	beforehand	physical match	3
Word	menu item	depend on candidate	name match	4
Image	image	none	class match	5 (max)

3.3.3 Selection Task and Interface

For gaze-button selection, we asked the participants to complete five selection tasks: *simple, key, word, icon*, and *image*. One trial involved completing a selection. Each task consisted of 51 trials. We used this number by considering the concentration and fatigue of the participants and used the first trial as a training trial (not used for our analysis). The order of the tasks among the participants was randomized. Before beginning the experiment, we calibrated the eye tracker with Tobii's 9-point calibration for each participant. The task began with the instruction display, which gave instructions to the participants for each task. The participants read the instructions and then pushed the space key to proceed. The task display was then shown, and the participants were asked to select a target using the gaze-button selection. Between the tasks, we asked the participants to take rest for at least one minute. The experiment took approximately 25 min. The task display included candidates specific to the task and one target. We did not give the participants visual feedback for all tasks to eliminate any potential side effects. We determined the target size at which the eye-tracking performance (i.e., the offset and precision) did not affect the selection, as described in each section describing tasks.

Although eye behaviors should be collected from various tasks, it is difficult to experiment with such diverse tasks. Therefore, we used these five tasks that represented daily interaction situations [IYS22]. We list the relationship between tasks and difficulties in Table 3.3.

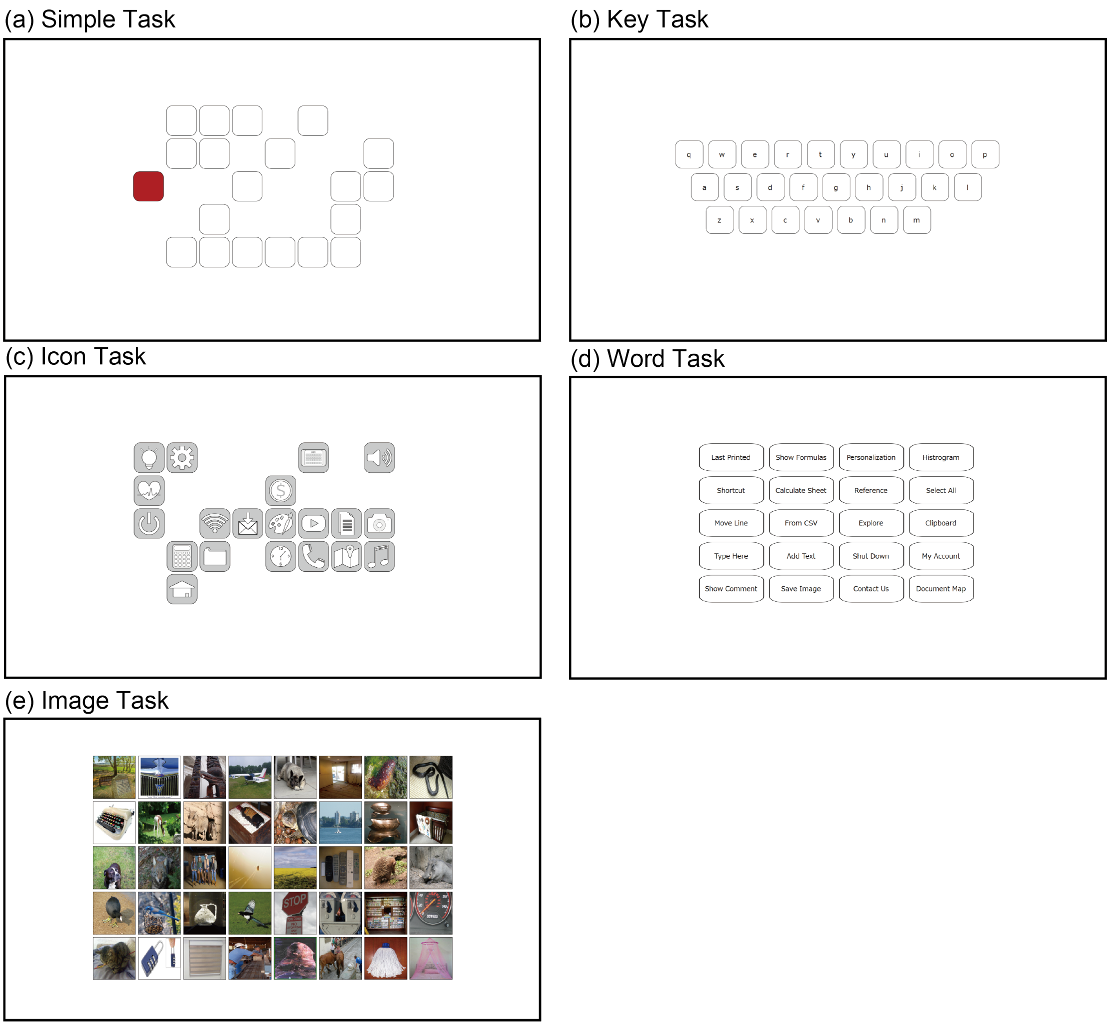

FIGURE 3.5: Displays for five selection tasks.

Simple Task

The simple task involves selecting a red rounded rectangle target (Figure 3.5a). We instructed the participants to "select a red object." This task is similar to the *simple reaction* task wherein a participant pushes a button after the visual stimulus is displayed [CNM83]. We displayed one red target and 19 white candidates in a random position in an 8×5 grid. The size of each target was 2.5°×2.5°.

Because there is one red target and the others are white, the participants need to not search for it and would know all candidates before the task. This selection corresponds to a real situation of a preprogrammed selection wherein users can select the target by just looking at it for a small duration. For example, a close button of the web browser can be selected by looking at it for a small duration. Such buttons are positioned at the top corner of the browser[2], and the user knows the content before looking at it. Selecting the most frequently selected targets is another real situation that this task imitates. Because these situations would be the easiest interaction situations, we defined the difficulty of the simple task as the lowest among the tasks.

Key Task

The key task involves selecting a key (Figure 3.5b). For example, we instructed the participants to "select [a] key." This task is similar to the *physical-match* task wherein a participant pushes a button if a visual shape of a candidate and an instruction are the same [CNM83]. We displayed 26 keys in a qwerty alignment. One of the 26 keys was randomly chosen as the target. The size of each target was 2.5°×2.5°.

Because we used a qwerty alignment, which the participants were familiar with, they had already known the position of all candidates and the content (i.e., a key). However, more recognition is needed to confirm a target than in the simple task. This task corresponds to a real situation of a key selection and a selection of a radio button with a character, such as selecting ⓐ, ⓑ, ⓒ, and ⓓ.

Icon Task

The icon task involves selecting a target that resembles a desktop icon (Figure 3.5c). For example, we instructed the participants to "select a [call] icon." This task is similar to the *name-match* task wherein a participant pushes a button if the meaning of a candidate and instruction are the same [CNM83]. We used an icon set comprising 20 icons that resemble desktop icons. As opposed to the key task,

[2]top-right in Microsoft Edge and top-left in Safari

the instruction and target differ (i.e., verbal instruction and visual target). We displayed one target and 19 candidates in a random position in an 8×5 grid. The size of each target was 2.5°×2.5°.

Before beginning the task, we asked the participants to memorize the correspondence between the images and instructions to eliminate preconceptions based on previous experience. The participants were required to recognize the object and then match the meanings of the object and instructions before pushing the button. This selection task corresponds to the real situation of a relatively simple image selection. For example, the desktop icons and tab-icons of the web browser whose position and image are already known by the user before looking at them.

Word Task

The word task involves selecting a one- or two-word target consisting of at least seven characters (Figure 3.5d). For example, we instructed the participants to "select a [copy text]." We created a word set comprising 20 words extracted from text- and image-editing interfaces such as Microsoft Word and Adobe PhotoShop. Similar to the key task, both the instruction and target are verbal. The only difference is the character length: one character vs. at least seven characters. We randomly selected one target and 19 candidates from the word set in a random position in a 4×5 grid. The size of each target was 5.5°×2.5°.

Unfortunately, there is no similar task in the MHP [CNM83]; however, the word task was used as the task that requires a higher cognition level than the one-character task [ZXZZ11]. Therefore, completing the word task is more difficult than the simple, key, and icon tasks.

Image Task

The image task involves selecting an image target (Figure 3.5e). For example, we instructed the participants to "select a [dog] icon." We used the image set extracted from Visual Genome[3]. Contrary to the icon task, we did not show all images to participants beforehand. While the icon and image tasks are both selection tasks against nonverbal candidates, there is a difference in whether the participants knew or did not know which images/icons were shown as candidates before a task has begun. This task is similar to the *class-match* task wherein a participant pushes a button if a class (e.g., a letter or digit) of a candidate and instruction are the same [CNM83]. We displayed one target and 39 candidates randomly selected from

[3] `https://visualgenome.org/`, licensed under CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`). (Retrieved October 13th, 2022)

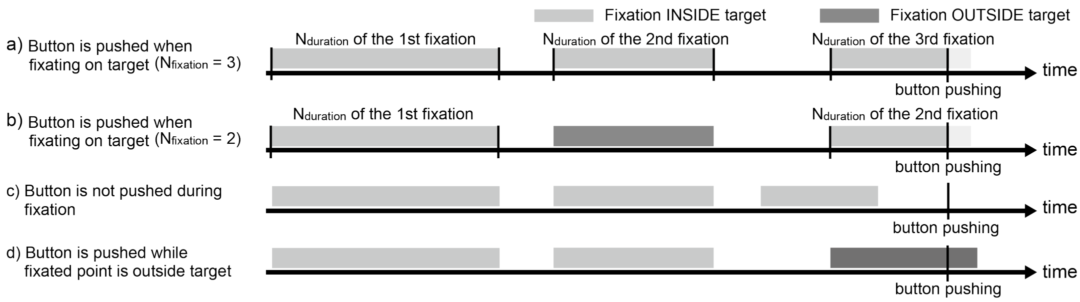

Figure 3.6: Fixations we used (a and b) and did not use for the subsequent analysis (c and d).

the image set in a random position in an 8×5 grid. The size of each target was 3.5°×3.5°.

The participants needed to recognize the object, classify it into an image type (e.g., an image of a dog), and match the classes of the object and instruction before pushing the button. This selection task corresponds to a real situation of relatively more complex image selection than that in the icon task. For instance, the images in an image-search result and an image that a user rarely sees. Therefore, the image task is the most difficult among all the tasks shown by Card et al. [CNM83].

3.3.4 Results

We measured $\mathrm{N}_{\mathrm{fixation}}$ and $\mathrm{D}_{\mathrm{fixation}}$ performed by the participant on a target before pushing the button. Accordingly, we validated our hypotheses and developed our model that derives the dwell time, which allows dwell selection to be performed after a user completes a cognitive process based on their behavior.

We discarded the first trial of each task as practice, and thus we used 1000 ($= (51-1)$ trials $\times$ 20 participants) trials for each task. Before detecting a fixation, we first excluded eye-tracking noise by applying the median filter with a window size of six samples, which is equal to 5 ms with 1200 Hz of the eye tracker. We then applied the I-DT algorithm [SG00] with a dispersion threshold of 30° and used 100 ms as the minimum duration of the fixation. Thus, in this analysis, the fixation consists of the gaze coordinates wherein the velocity of gaze movement is below 30°/s over 100 ms. We used specific fixations wherein the fixation point (i.e., the centroids of gaze coordinates during the fixation) was inside the target. Furthermore, we used trials wherein the participant pushed a button and successfully selected during fixation (Figure 3.6a and b), making trials consistent in the analysis. We did not use the trials wherein selection was not done during a fixation (Figure 3.6c), and the fixation was outside a target (Figure 3.6d). This process of

TABLE 3.4: N_{fixation} required for completing each task. The number in the parentheses is that of the participants. For example, twelve participants required two fixations in 20 trials to complete the key task.

Task \ N_{fixation}	1	2	3	4	5	6	7
simple	987 (20)	3 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
key	946 (19)	20 (12)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
icon	865 (20)	87 (20)	6 (6)	5 (5)	0 (0)	0 (0)	0 (0)
word	768 (20)	172 (20)	23 (14)	3 (3)	1 (1)	0 (0)	0 (0)
image	548 (20)	308 (20)	79 (20)	3 (3)	2 (2)	2 (2)	0 (0)

fixation detection was necessary owing to the eye-tracking noise and our definition of fixation. For example, some noise may have remained and been affected by the algorithm. Given that we did not use the trials wherein selection was not made during a fixation (Figure 3.6c), these trials were determined as errors, although the participant successfully selected a target.

Number of Fixations

We show the N_{fixation} that the participants are required to complete each task in Table 3.4. In total, we detected fixations for 4,828 trials; we could not detect fixations in 172 trials (3.4% of all trials). We did not instruct participants on the selection strategy (e.g., select a target with a small N_{fixation}) to observe participants' natural selection behavior. Although the participants did not frequently require a large N_{fixation}, they seemed to require it (e.g., $N_{\text{fixation}} \geq 3$) for completing tasks with high difficulties (i.e., icon, word, and image tasks). Thus, we concluded that this result verifies **H1** that *N_{fixation} required for selecting a target increases along with the difficulty of our task.*

Duration of Fixation

We first measured the D_{fixation} of the last fixation (i.e., fixation when the target was selected). Because the last fixation included the participant's button pushing in our analysis, we used the D_{fixation} of the last fixation as the duration required for recognizing the target and making a decision thereafter. We show the average D_{fixation}s of the last fixation for each N_{fixation} required for completing a trial in Figure 3.7. The average D_{fixation} tends to decrease N_{fixation} increases. When N_{fixation}

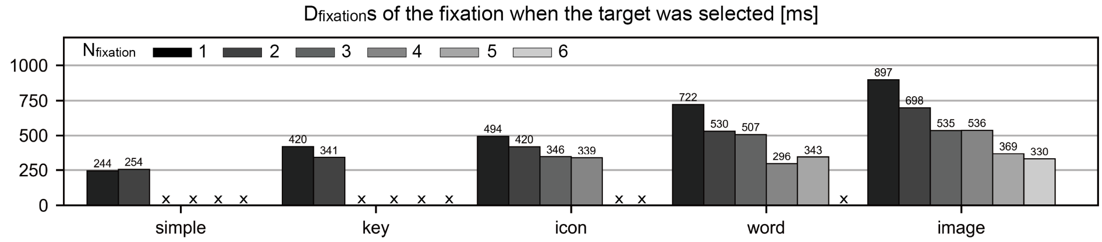

FIGURE 3.7: $D_{fixation}$ of the last fixation for each task and each $N_{fixation}$. For example, for image task, $D_{fixation}$ averaged in 330 ms for $N_{fixation}$ was six (i.e., the participants required six times of fixations to complete the trial.

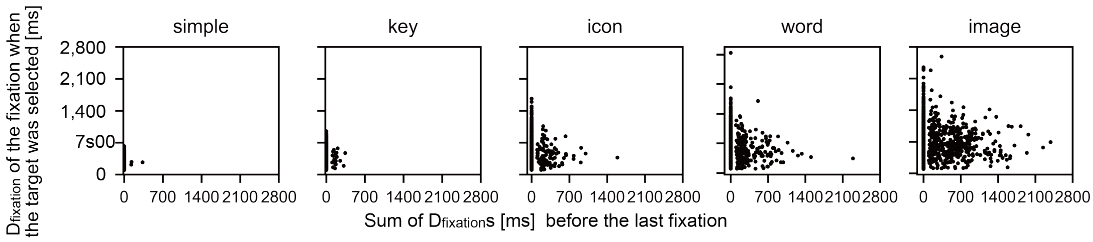

FIGURE 3.8: $D_{fixation}$ of last fixation against the sum of $D_{fixation}$s before last fixation.

TABLE 3.5: Regression results for each task on a linear model of Equation 3.2: $a + b \times (\mathrm{N_{fixation}}\text{-}1)$.

Simple	Key	Icon	Word	Image
R^2 =1.0	R^2=1.0	R^2=0.918	R^2=0.860	R^2=0.935
a =244.2	a = 420.4	a = 480.5	a = 677.7	a = 834.1
b =10.2	b = -79.4	b = -53.9	b = -99.0	b = -109.2
AIC =-111.8	AIC = -110.5	AIC = 38.5	AIC = 58.5	AIC = 67.8

TABLE 3.6: Regression results for each task on a logarithmic model of Equation 3.3: $a + b \times \log_2(\mathrm{N_{fixation}})$.

Simple	Key	Icon	Word	Image
R^2=1.0	R^2=1.0	R^2=0.972	R^2=0.907	R^2=0.972
a=244.2	a = 420.4	a = 494.3	a = 721.9	a = 905.8
b=10.2	b = -79.4	b = -82.6	b = -175.4	b = -217.9
AIC= -111.8	AIC = -110.5	AIC = 34.1	AIC = 56.5	AIC = 62.8

was one (i.e., the participant fixated a target once), $\mathrm{D_{fixation}}$ increased as the difficulty of the tasks increased. We then investigated the relation between the $\mathrm{D_{fixation}}$ of the last fixation and the sum of $\mathrm{D_{fixation}}$s before the last fixation (Figure 3.8). For example, if $\mathrm{N_{fixation}}$ is three, we calculate the sum of the first two $\mathrm{D_{fixation}}$s; and if $\mathrm{N_{fixation}}$ is one, the sum becomes zero. This relation indicates that the $\mathrm{D_{fixation}}$ of the last fixation decreases as the sum increases, indicating that when the participant fixated on a target for a long time, they could make a decision in a small duration. Although certain $\mathrm{D_{fixation}}$s did not decrease as the difficulty increased, and the sum of $\mathrm{D_{fixation}}$s before the last fixation (e.g., between one and two $\mathrm{N_{fixation}}$ for the simple task as shown in Figure 3.7), these results may verify **H2** *D_{fixation} of the fixation, when the target is selected, decreases as total N_{fixation} increases.*

3.4 Model Deriving Dwell Time

In this section, we show the model using the results of the experiment and the MHP. Then, we explain how we developed the model by validating the three hypotheses.

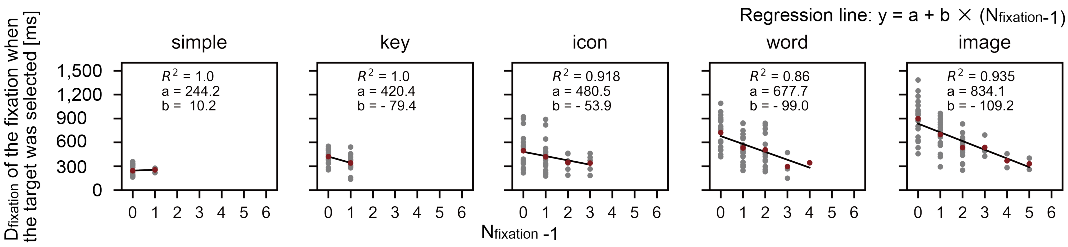

FIGURE 3.9: Regression results with average D_{fixation} and linear equation (Equation 3.2) for each task. Gray plots are average D_{fixation} for each participant. Red plots are average for each N_{fixation}.

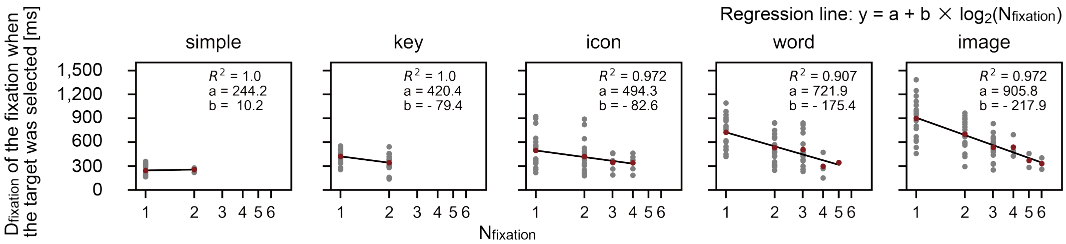

FIGURE 3.10: Regression results with average D_{fixation} and logarithmic equation (Equation 3.3) for each task. Gray plots are average D_{fixation} for each participant. Red plots are average for each N_{fixation}.

3.4.1 Equations of Model

To evaluate our model, we first examined where N_{fixation} linearly affects the duration using the following equation:

$$y = a + b \times (N_{\text{fixation}} - 1). \tag{3.2}$$

We then explored the following equation as a more precise model wherein N_{fixation} logarithmically affects the duration:

$$y = a + b \times \log_2(N_{\text{fixation}}). \tag{3.3}$$

In these two models, y indicates the duration in a certain N_{fixation}, a indicates the duration when N_{fixation} is one, and b indicates a change in the D_{fixation} of the last fixation as N_{fixation} increases.

We show the regression results of the linear model in Table 3.5 and Figure 3.9 and the logarithmic model in Table 3.6 and Figure 3.10. The R^2 in Equation 3.3 was higher than that in Equation 3.2 for the icon, word, and image tasks. Because the maximum N_{fixation} was two in the simple and key tasks, R^2 was 1.0. Because a human can remember a stimulus (visual image in this work) and proceed with the processes in cognition subsystems by referring to the preprocessed stimulus [ZXZZ11], the time required for the processes seemed to decrease as N_{fixation} increased. We assumed that this is what caused a higher R^2 in Equation 3.3

Further, we compared the *AIC* values [Aka74] of the two models to determine an appropriate model statistically. As a brief guideline, a model with a lower *AIC* is better, and a model with $AIC \leq (AIC_{\text{minimum}} + 2)$ is probably comparable with better models [BA03]. Thus, we determined to use $\log_2(N_{\text{fixation}})$ as an independent variable of the expression in our model.

3.4.2 Slope in Our Model

To interpret slope b based on the relation that was derived from the regression result and the slope estimated by MHP, we justify **H3**. The slopes of the equations (i.e., b in Equation 3.3) show a downward trend from 10.2 (simple task) to −217.9 (image task) as the difficulty of the task increases. We compare the slopes and the estimated slopes using the MHP, as shown in Table 3.7. The estimated slopes using MHP are 0 ms ($0\tau_c$), 70 ms ($1\tau_c$), 140 ms ($2\tau_c$), and 210 ms ($3\tau_c$) in simple reaction, physical match, name match, and class match tasks, respectively. The differences between the slope with our model and the estimated slope are under 35.4 ms. Because the original τ_c also ranged from 25 ms to 170 ms, this difference could be considered to be covered by the range. Thus, we can estimate slope b from the number of required processes of *recognize, classify*, and *match* multiplying by τ_c (70 ms).

3.4.3 Minimum D_{fixation} for Each Task

We investigate the minimum D_{fixation}, that is, D_{fixation} for that N_{fixation} is the largest considering that we showed the D_{fixation} of the last fixation decreased as N_{fixation} increased in Section 3.3.4[4]. The minimum D_{fixation} was 244.2 ms (N_{fixation}=1) for the simple task, 341.0 ms (N_{fixation}=2) for the key task, 329.3 ms (N_{fixation}=4) for the icon task, 314.6 ms (N_{fixation}=5) for the word task, and 342.5 ms (N_{fixation}=6) for the image task. There is a difference of approximately 90 ms between the simple task (244.2 ms) and other tasks (331.9 ms on average) owing to the fact that the simple task requires only *requesting*, while the others require at least one process of *recognize, classify*, and *match* in addition to *requesting*. Moreover, the value of 90 ms is within the range of τ_c (25–170 ms). Therefore, we concluded that the

[4]except for the simple reaction task.

TABLE 3.7: Relation between slope and estimated slope using MHP. Units of all digits are in milliseconds. The estimated slope was calculated with the number of required processes *recognizing*, *classifying*, *matching* for each task. For example, in image task, four processes of *recognizing*, *classifying*, *matching*, and *requesting*, require $4\tau_c$. Because a *request* process can be regarded as the process of decision-making for tasks described in [CNM83], we exclude τ_c for *request* from one cognitive cycle, the estimated slope with MHP is similar to $3\tau_c$= 210 ms.

Task	Required cognitive process				Slope (ours)	Slope (MHP)	Diff.
Simple				*request*	10.2	0.0 ($0\tau_c$)	10.2
Key			*match*	*request*	−79.4	−70 ($1\tau_c$)	9.4
Icon			*match*	*request*	−82.6	−70 ($1\tau_c$)	12.6
Word	*recognize*		*match*	*request*	−175.4	−140 ($2\tau_c$)	35.4
Image	*recognize*	*classify*	*match*	*request*	−217.9	−210 ($3\tau_c$)	7.9

difference in the D_{fixation} of the last fixation between the simple task and other tasks could be interpreted due to the difference in the required processes for decision-making.

Because users can generally select the target in a well-familiarized interface without careful fixation, even if the target is a key, icon, word, or image, there is a possibility of selecting a target without a cognitive process. In other words, there is a possibility that they can make a decision as being equal to the simple task. For example, because users who are familiar with the current interface in Windows and MacOS know that the home icons are often located in the corner, they can potentially select the icon without careful fixation. Thus, we concluded that one minimum D_{fixation} exists regardless of the task and D_{fixation} converges to the one in the simple task (i.e., 244 ms), which verifies **H3**: *D_{fixation} for large N_{fixation} converges to the duration required for completing decision-making processes for a simple reaction task regardless of the task.*

3.4.4 Range of D_{fixation}

In addition to the aforementioned analysis focusing on average values, we analyzed how the D_{fixation} in each N_{fixation} varied among participants (Figure 3.10). These ranges may be attributed to the same factor as in the MHP, that is, the Fastman can complete a task with minimum duration, and the Slowman requires maximum duration. Because we did not instruct participants on the selection strategy, D_{fixation} also varied for each participant and selection. Personality and background may

TABLE 3.8: Summary of regression results for each task.

Task	Equation	Max N_{fixation}	Smallest dwell time
Simple	$174.2 + 10.2 \times \log_2(N_{\text{fixation}})$	2	174.2
Key	$350.4 - 79.4 \times \log_2(N_{\text{fixation}})$	2	271.0
Icon	$424.3 - 82.6 \times \log_2(N_{\text{fixation}})$	4	259.3
Word	$651.9 - 175.4 \times \log_2(N_{\text{fixation}})$	5	244.6
Image	$835.8 - 217.9 \times \log_2(N_{\text{fixation}})$	6	272.5

have also affected the results. For example, a user carefully searching for a target requires a large τ_c, and a user familiar with a target (e.g., a user has used the menu item in the word task) requires a small τ_c. Therefore, using average values is a generally simple solution to reflect the duration that a human requires for the decision-making process. However, using a calibrated τ_c for users is a better solution to estimate a more precise duration.

3.5 Applying Our Model for Dwell Selection

In this section, we describe how our model can be applied to dwell selection. Because no action of pushing a button is required for dwell selection, we first subtract a duration of τ_m=70 ms from the model. By using Equation 3.3 and the regression results, we define the adapted model for each task. We summarize the equations and dwell times derived using our model for each task in Table 3.8, which indicates that we can dynamically change dwell times using our model. For example, in an image-selection task, if a user fixates on a target three times beforehand, we can use 490.4 ms as the dwell time (=835.8 - 217.9 $\times$ $\log_2(3)$); if six times, we can use 272.5 ms (=835.8 - 217.9 $\times$ $\log_2(6)$).

We consider a span that keeps counting N_{fixation}. First, our idea is to use average durations for the trial (i.e., from displaying a target to finishing a selection) in the experiment as the span; the duration was 609, 996, 2,455, 3,620, and 23,565 ms for the simple, key, icon, word, and image tasks, respectively. For example, for the task of selecting an image, the system keeps counting N_{fixation} during 23,565 ms and calculates the dwell time with the counted N_{fixation}. We did not consider N_{fixation} more than those observed in our experiment (more than max N_{fixation} in Table 3.8) and determined the minimum N_{fixation} for each task. However, as described in Section 3.4.3, the minimum D_{fixation} may become one for the simple task (i.e., 174.2 ms). Of course, if users prefer a faster interaction, they can use under 174.2 ms at will. Such a small dwell time can be considered when users are familiar with the situation.

Although we have described the use of our model in a real interaction, we cannot strongly conclude that it is useful mainly owing to the limitations of our experimental conditions and results. Therefore, further investigation with an application adopting dwell selection with our model should be conducted.

3.6 Conclusions

In this chapter, we developed a model that derives the dwell time, which enables dwell selection after a user completes the decision-making process based on their eye behavior.

We first conducted an experiment involving five tasks of different difficulties to measure the number of fixations and their duration based on the eye behaviors of participants during the selection task. We then validated our three hypotheses related to fixations and developed our model using the fixations and durations by referring to the MHP. Then, we demonstrated how our model derives the dwell time.

We positioned this work as a first step work to answering RQ1: *How should we determine dwell time?* more deeply. The results showed that the dwell time can be determined using a fixation behavior that users subconsciously did for completing selection tasks and knowledge of decision-making processes. Based on these findings, we showed that we could determine dwell times that answered the question.

However, our model is not the only model to answer the question; there are limitations and huge design space for developing a dwell time determination method. First, our findings are limited by the experimental conditions. It is unclear whether our findings, i.e., the duration that a human requires to finish a cognitive process, would hold under other conditions. Regarding the selection tasks, there are numerous situations of real interactions, for example, selecting a thumbnail, which comprises an image and sentence and object of a movie. Second, because τ_p, τ_c, and τ_m were derived from certain user attributes [CNM83], our model may not be suitable for users whose attributes differ from those of the participants in this experiment (e.g., different ages, experience with computer interaction, and experience with gaze-based interaction). However, this is only a hypothesis, and we could not make specific conclusions from our current results; therefore, further investigation for a large number of participants and more diverse participants is required. Although we concluded that our model based on Equation 3.3 could effectively derive duration, it is necessary to evaluate the model under other experimental conditions.

We developed our model from the perspectives of linear- and logarithmic-based equations and the MHP [CNM83]. Similar to Fitts' Law [Fit54] and ACT-R [AMD95, AML97], which has numerous variations of a model regarding the context, we can explore a variation of our model for a specific context or user attributes. For example, the keystroke-level model [CMN80] indicates that the time to complete a typing (i.e., key selection) task varies depending on the context and the user's typing skill. Similar to previous studies on adjusting dwell time (e.g., [MWWM17]), our model can be improved using the keystroke-level model. Similarly, we used the MHP to interpret human decision-making processes; however, there are numerous models for interpreting human cognitive processes (not limited to the decision-making processes), as discussed above. Therefore, we should further consider and compare our model with various models for the development of a more accurate and plausible dwell time determination method.

Chapter 4

USER INTENT DETECTION WITH MULTIPLE NATURAL EYE BEHAVIORS FOR DWELL SELECTION

In this chapter, we present a model that detects user intents to interact with a computer, especially for selecting a GUI object, by incorporating multiple natural human eye behaviors. We then apply the model to dwell selection to solve Midas-touch, which is a long-term issue in gaze-based interaction.

We use eye movement, saccade, fixation, pupil diameter, and vergence as eye behaviors for intent detection, which can be calculated from the data sampled by the eye tracker. These eye behaviors may involve user attention and intent, and hence they have been used in various studies [DJPZ+21, BVH12, SA00]. Because eye behaviors generally rely on users, ambient environment, and interaction situations, identifying which eye behaviors and their characteristics are useful to interpret user intent can be challenging. For example, it is difficult to interpret user intent from threshold-based methods, such as "if the pupil diameter enlarges over 1 mm, that behavior indicates user intent to interact" because the pupil diameter generally depends on the ambient light. Therefore, we adopt a machine learning (ML)-based method to interpret user intent from these eye behaviors. We do not focus on each behavior in detail but focus on the possible features calculated from those eye behaviors as cues of user intent. We collect the eye behaviors from five different tasks to investigate how the eye behavior differs among tasks and attempt to develop a general ML model for users and tasks.

We first introduce the overview of our dwell selection (i.e., DTD-ML selection), investigate natural human eye behavior during the selection task, develop a user intent detection model using the obtained eye behaviors, and then evaluate the performance of our dwell selection.

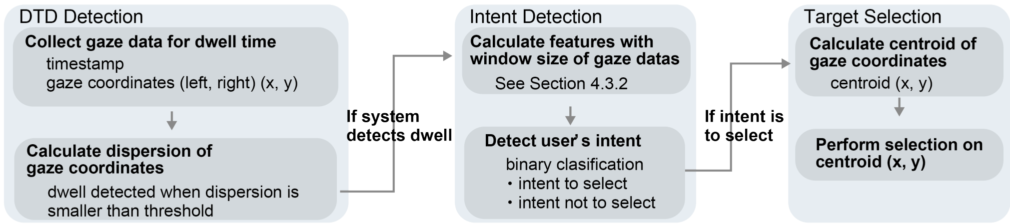

FIGURE 4.1: Overview of the DTD-ML selection system.

The contributions of this work are as follows.

- We develop a user intent detection model based on ML by incorporating multiple natural human eye behavior and apply the model for dwell selection (DTD-ML selection).
- We collect labels for creating an ML-based intent detection model from five different tasks, representing four interactive situations and one everyday situation without manipulation.
- We show that our intent detection model achieves an area under the curve (AUC) of the receiver operator characteristic (ROC) curve of 0.903; it also achieves high AUC values independent of the user and eye-tracking frequency, as described in Section 4.3.
- We show that the DTD-ML can prevent 40.2% of unwanted selections compared to DTD selection and has equal or better usability than both the DT and DTD selection methods.

4.1 Our Dwell Selections

Figure 4.1 shows how a system detects user intent, either to select or not to select and triggers selection. Our system comprises three parts: DTD-based user intent detection (DTD detection), ML-based user intent detection, and selection.

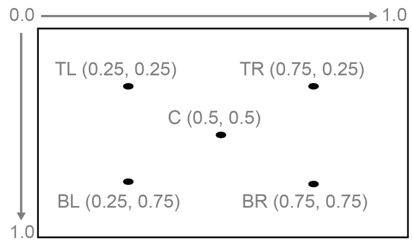

FIGURE 4.2: The display used for investigating dispersion in the preliminary experiment to determine the dispersion threshold. Points were instructed points where participants looked.

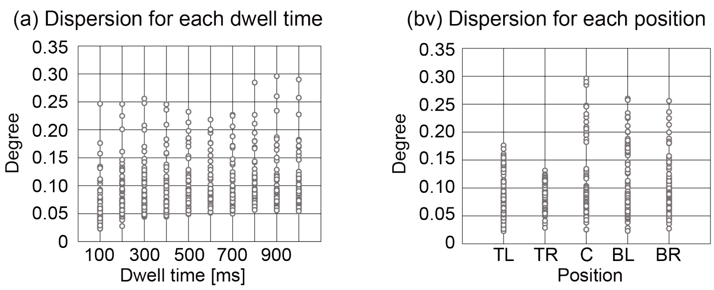

FIGURE 4.3: Results of the preliminary experiment to determine dispersion threshold: dispersion results for each dwell time (a) and position (b).

4.1.1 Dwell Time and Dispersion (DTD) Based User Intent Detection (DTD Detection)

In our system, DTD-based user intent detection contributes to a rough screening of the user's intent to select and trigger ML-based intent detection. The DTD detection system detects a dwell if the dispersion during the dwell time is less than a dispersion threshold. Owing to the dispersion threshold, the user needs to dwell more intentionally than in DT selection. However, this helps prevent the Midas-touch problem.

We determined the dwell time and dispersion threshold from a preliminary investigation considering there is no detailed investigation of suitable thresholds, although DTD detection has been used in commercial software [HWM+89, SJ00, SRT11, TA08] and for other interactions [ULH10, IYS20, HC05, HCH04, KMS10, Dyn21]. In particular, we investigated the dispersion in the user's gaze in a certain dwell time while intentionally dwelling on a point. Fourteen male volunteers (aged 21–25) participated in this investigation. 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 at a distance of approximately 65 cm from the display. We asked the participants to calibrate the eye tracker before starting the first task. Participants looked at each of the five points on display, as shown in Figure 4.2, for 2000 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°/s [SG00]. To obtain 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 gaze coordinates in the visual degree for 10 dwell times (100, 200, 300, ..., 900, 1000 ms; if the dwell time is 100 ms, we used the gaze coordinates in the visual degree of the first 100 ms (i.e., 1,000 ms to 1,100 ms out of 2000 ms.)) as the dispersion. The results showed that the dwell time did not affect the dispersion. Moreover, all dispersions were less than 0.3° regardless of the dwell time and the position (Figure 4.3). Therefore, we used 0.3° as the dispersion threshold and 600 ms as the dwell time in our system. This study chose 600 ms as dwell time for two reasons; first, it is not a large dwell time compared to those in previous studies; second, it is an appropriate dwell time for the cognition model [IYS21][1]. 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 should be addressed as future work.

[1]The work in Chapter 4 is done before the work in Chapter 3. Thus, we did not use the dwell time by using our model in Chapter 3.

4.1.2 Intent Detection with an ML Model

After dwell detection, the system detects the intent either to select or not select (i.e., binary classification task) using the ML model. The system first calculates features from the window size (2000 ms in this work, as described in Section 4.3.4) of the gaze data before a dwell is detected. For the features, we use eye behaviors of saccades, fixations, vergences, pupil changes, and quantitative data of eye movement distances and durations, which have been described in detail in Section 4.3.2.

4.1.3 Target Selection

If the detected intent is to select, we calculate the centroid of the gaze coordinates, $C_{x/y}$, during the dwell. The system then activates selection to $C_{x/y}$.

4.2 Experiment 1: Labeling of User Intent

In this experiment, we collected ground-truth labels that represent the user intent to either select or not select.

4.2.1 Participants and Apparatus

We recruited 24 university students (five females and 19 males, all Japanese) aged 20–26 (M = 22.9). 15 participated in the 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. We used two different eye trackers since it was necessary to investigate whether we could use our ML model with different eye trackers, considering the eye-tracking frequency generally differs from one device to another. Additionally, some participants could not calibrate the Tobii Pro Fusion due to the incompatibility of pupil detection for Asians[2].

12 participants used the Tobii Pro Spectrum at 1200 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

[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.

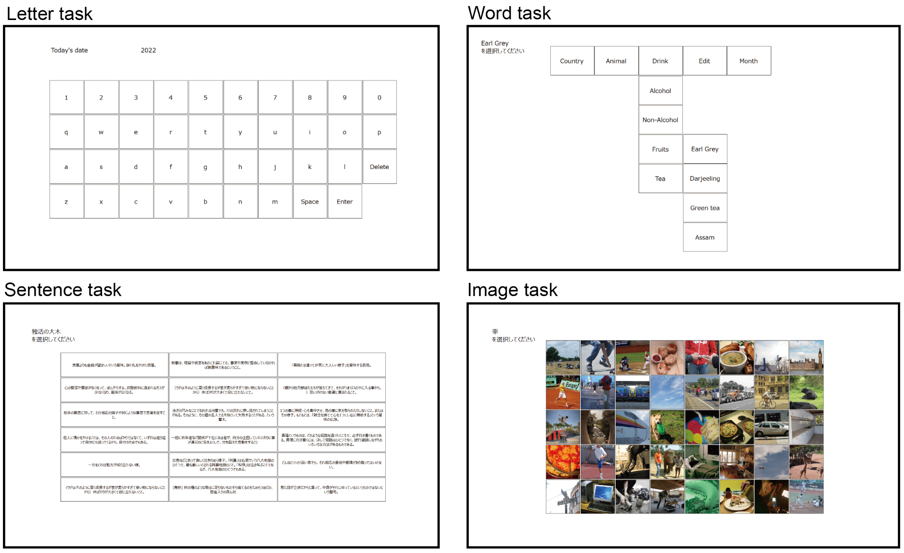

FIGURE 4.4: Displays used for tasks.

experimental environment is shown in Figure 3.4. The experiment was conducted in a room with fluorescent light at approximately 810 lux.

4.2.2 Tasks

Because eye behaviors vary according to the task, environment, and visual stimulus, the experiment should be conducted in diverse conditions. However, it is difficult to incorporate all of the diverse conditions. In this work, we collected labels and eye behaviors from five different tasks: a letter task, a word task, a sentence task, an image task, and a movie task, which 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 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.

Figure 4.4 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 the selection. The participants read the task instruction and pushed the space key to move on. Regardless of the participant's intent, the system displayed the labeling form, which contained only a questionnaire regarding the intent when it detected a dwell. To eliminate any potential side effects, we did not give the participants visual feedback for all tasks; however, they could recognize that a dwell was detected through the labeling form's appearance, except in the movie task.

Letter Task

The participants successively selected keys on a displayed keyboard. The size of each key was $3.5° \times 3.5°$. The keyboard comprised 10 digits, qwerty-arranged keys, a space key, a delete key, and an enter key. The task involved typing the date and the participant's name, age, and hobby; e.g., one instruction was "write today's date." For example, for the instruction "write today's date," the participants typed the date using the displayed keyboard. There was no specific format, and thus, the participants could enter the date freely, e.g., "20210801" or "0801." They finished the trial by selecting the enter key and labeling their intent for each key selection.

We assume that this task represents a situation wherein the user selects a letter; a selection of a radio button with a character, such as selecting ⓐ, ⓑ, ⓒ, and ⓓ is another possible situation.

Word Task

The participants manipulated a three-layer hierarchical menu and selected an item that was written in word(s). The size of each item was $4.5° \times 3.0°$. 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 the target as appropriately as possible; if they could not find one, we asked them to select an arbitrary target. We did not limit the number of times the menu could be opened 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 is another possible situation.

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 $11.0° \times 2.5°$. Each participant performed 30 selections for randomly chosen phrases. From the 100 pairs, we arranged 18 choices, comprising one correct meaning and 17 randomly chosen meanings, in a 3×6 grid. We asked the participants to select the choice as correctly as possible; if they could not find one 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 wherein the user reads sentences and selects one, such as a hyperlink on a web page.

Image Task

We asked the participants to select an image that was appropriate for a given verbal instruction. We used a set of 64332 images[4], and the size of each image was $3.5° \times 3.5°$. Each participant performed 100 selections for randomly chosen instructions. We arranged 40 choices comprising at least one correct image along with randomly chosen images in an 8×5 grid. We asked the participants to select the image as correctly as possible; however, 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 wherein 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.

Movie Task

Contrary to the other tasks, we informed 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 [FCHN15] and streamed them using a full-screen mode of Windows Media Player without a UI. Each participant watched movies for 10 min. They were allowed to be absent-minded if a movie was not attractive.

Although this task involves simply watching a movie on a desktop computer, the gaze data collected through it involves various kinds of information. For example, because we chose the movies regardless of the participants' interests, we could

[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 4.1: Guidelines for intent labeling.

Situation wherein 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 the form was displayed but participant still thinking about target's correctness	No (negative class)
Participant thinking, searching, or lost in thought	No (negative class)

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. Therefore, we conducted this movie task to collect negative data representing gaze data that did not involve intentional manipulation in daily life. In the movie task, we did not instruct participants to label their intent.

Intent labeling

The participants gave their intent concerning dwell detection with a physical keyboard following the guidelines listed in Table 4.1. In the following analysis, we used the detected dwells that were labeled "Yes" and "No" as the positive and negative classes, respectively. The selection labeled as the negative class was treated as an unwanted selection. Note that there was no selectable UI for the movie task, and the participants did not label their intent; accordingly, we labeled all the detected dwells in the movie task as negative classes.

4.2.3 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 min per participant.

4.2.4 Labeling Results

The labeling results are summarized in Table 4.2. Even without the ML-based intent detection, there were fewer negative classes for the letter task than for the other tasks owing to the fact that the letter task possesses a lower cognitive load than the others. Contrary to the letter task, participants said that the sentence task was challenging because it was difficult to find the correct meaning of the

TABLE 4.2: Numbers of labeled classes.

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

Japanese phrase. This indicates that they spent much time reading and thinking about the sentence, resulting in more Midas-touches. Therefore, the negative class percentage was the highest for the sentence task. For the movie task, there is a total of 10586 negative classes, which suggests that there are many possibilities for the mis-detection of user intent and occurrence of Midas-touch during everyday situations.

Note that although balancing the number of labels for each task is preferable, 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 ignore data collected from the letter task, whose labels were biased positively, the detection may result in false negatives. In contrast, if we ignore data collected from the movie task, whose labels were biased negatively, the detection may result in 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.

4.3 Model Detecting User Intent for Dwell Selection

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

4.3.1 Data Processing

As listed in Table 4.2, 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. In particular, to achieve a 50:50 ratio, we randomly chose negative classes from the movie task for each participant.

To calculate the features, we used 2000 ms as the window size of the gaze data before a dwell was detected; a detailed explanation is given in Section 4.3.4. The

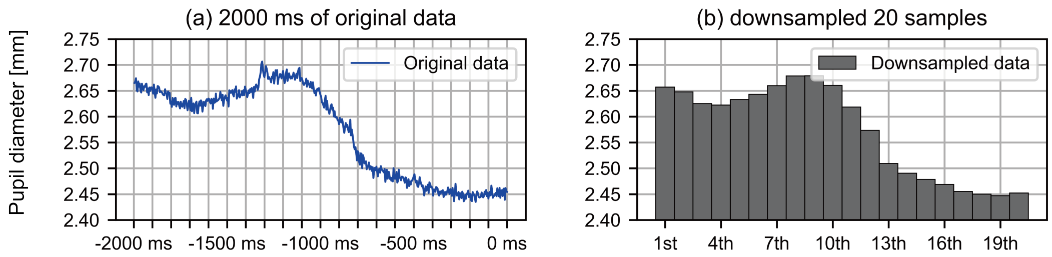

Figure 4.5: Example of raw data for (a) **pupil** and (b) its downsampled values.

gaze data were the x/y coordinates ([0.0 (top left) –1.0 (bottom right)]) on 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 the gaze data. Figure 4.5 shows an example for **pupil**. Next, we calculated the relative values between the last (20_{th}) value and each i-th ($i = 1, 3, ..., 19$) value (19 changes). Based on the changes (instead of the original values), we eliminated 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 1200 Hz) because gaze data do not change within a short span [Cly62], and eye-tracking data contains noise. Moreover, we adopted 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 [SG00] to detect fixations and saccades with the original x/y coordinates. For the parameter of the I-VT algorithm, we used 10°/s for fixation detection and 100°/s for saccade detection. Moreover, to exclude eye-tracking noise, we used 100 ms as the minimum duration of fixation and 30 ms as the minimum duration of a saccade.

4.3.2 Features

We used the following gaze data to calculate the features that are listed in Table 4.3.

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.

TABLE 4.3: 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**, $\mathbf{diff_x}$, 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**, $\mathbf{diff_x}$, and **pupil**	1st value, 19th value, difference between 19th and 1st values	12 (4×3)

$\mathbf{diff_x}$**:** 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.

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 4.3 are consistent with those used in previous works [BVH12, DJPZ+21]. Because these statistical values summarize the original data and would allow the detection model to focus on the important characteristics, the detection result may be better than using the original data. In general, the directions of the changes 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 for each sign and with both 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 difficult to decide the thresholds for each feature. We thus use ML-based detection.

Table 4.4: Summary of our intent detection. Values except for *all* are average values. MCC means Matthews correlation coefficient. We highlighted important results with aspects of contribution (red) and limitation (blue).

	AUC	accuracy	recall	precision	F1	MCC
all	0.903	0.826	0.839	0.818	0.828	0.652
all (hyper-parameters)	0.905	0.829	0.845	0.819	0.832	0.659
each-participant	0.893	0.819	0.831	0.817	0.822	0.64
each-task	0.964	0.952	0.965	0.972	0.968	0.746
each-frequency	0.909	0.835	0.85	0.827	0.838	0.67
leave-one-participant-out	0.898	0.812	0.828	0.81	0.812	0.634
leave-one-task-out	0.601	0.689	0.721	0.853	0.778	0.084
leave-one-frequency-out	0.880	0.793	0.808	0.79	0.792	0.595

4.3.3 Metrics for Evaluation

We used the area under the curve (AUC) of the receiver operating characteristic curve (ROC) [Bra97] as the primary metric for evaluating the detection performance. A higher AUC value indicates a greater chance of achieving both a high true positive rate (TPR) and a high true negative rate (TNR), and this helps our detection system deal with the trade-off between the Midas-touch problem and the ease of selection.

4.3.4 Creating ML Model

We created detection 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 model.

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 4.3.4).

Overall detection Results

Table 4.4 summarizes the detection 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 detection probability threshold. The curves of TPR and TNR 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, similar to the dwell time, there is a trade-off between the TPR and TNR.

We also provide the detection results obtained using hyper-parameters that we determined by using LightGBM Tuner from Optuna [ASY+19]. 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.

Detection 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 detect the user's intent and can thus be used as a general model independent of the user. Given the limited diversity of the participants, their small age range may have resulted in high AUC values. However, because we did not use the original values for **pupil** and $\mathbf{diff_x}$, which vary according to individual, as features, similar results may be achievable for users with different attributes.

Detection 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 detections, especially for the sentence task, whose AUC value was 0.443.

Because the movie task had one class, we did not create a detection 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 these four tasks to equalize the class ratio. The testing yielded a TNR of 0.463 when the detection 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* highlight the significance of using more various tasks when creating a gaze-based intent detection model.

Detection 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 possessing different frequencies. However, eye trackers mounted on an HMD and different eye-tracking (or pupil-tracking) methods may yield different results.

Detection Results for Window Size

We examined the detection 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 were shorter than 3,000 ms. Thus, when we used 3,000 ms as the window size, approximately 20% of the tasks were eliminated compared to when 600 ms was used 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. Based on these results, we created features using a window size of 2,000 ms, which was the smallest one that achieved an AUC value greater than 0.900.

Detection Results for Other Classifiers

We examined the detection 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 previously.

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 [HP60]. While the use of those values increases the AUC values for *all* (>0.940), they may

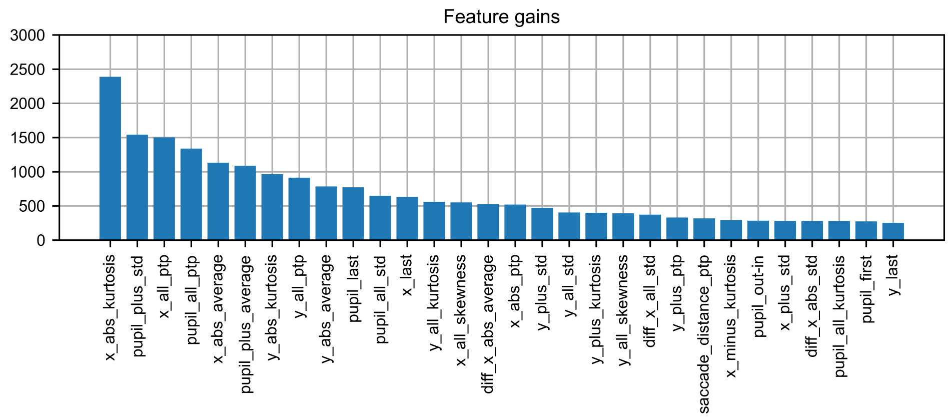

FIGURE 4.6: Feature importance for our intent detection method.

have caused overfitting that could not be displayed in the detection test with the current data.

Feature Importance

The top 10 gains among the features were the average and kurtosis of the absolute values of the **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**, as shown in Figure 4.6. This result suggests the significance 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 significant impact because the diameter increases with interest or emotion [HP60].

4.4 Experiment 2: Performance Evaluation

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

4.4.1 Participants and Apparatus

We recruited 12 university students (four females and eight males, all Japanese) aged 20–24 (M =22.9). Six participated in Experiment 1 and nine participated in an experiment with a gaze-based interface. This experiment used the same

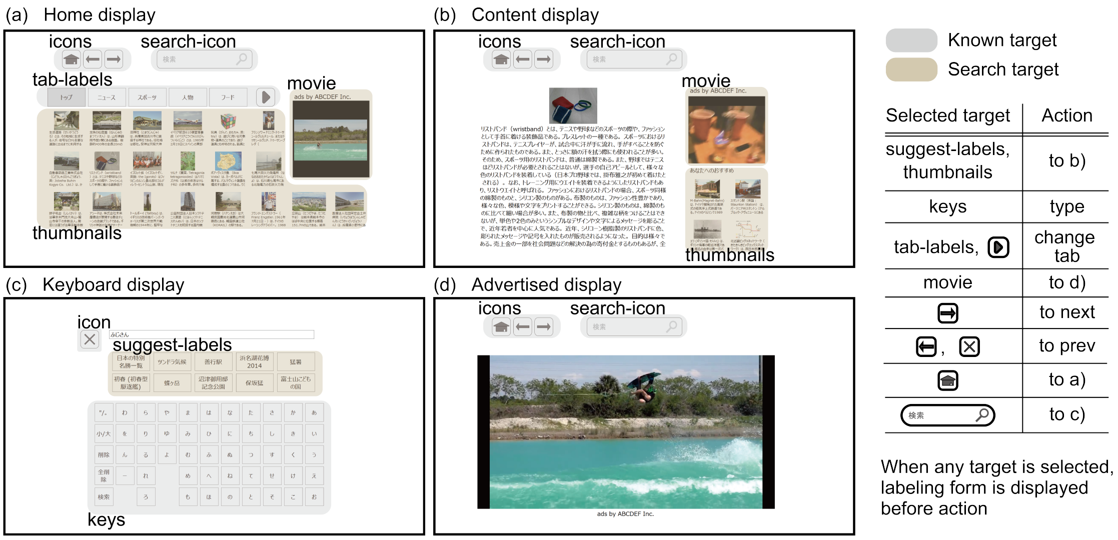

FIGURE 4.7: Interface used in Experiment 2.

apparatus and environment as Experiment 1. In particular, we used the Tobii Pro Spectrum at 1200 Hz as the eye tracker in Experiment 2.

4.4.2 Task

The task was to interact with a dictionary-like interface, shown in Figure 4.7, using dwell selection. We roughly classified the targets in the interface into two types: the *known target*, wherein the participants knew the location and content, and the *search target*, wherein 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 content 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 that used 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 were $2.0° \times 2.0°$ for keys and icons, $4.0° \times 2.0°$ for tab-labels and suggest-labels, $4.0° \times 4.0°$ for thumbnails, $8.0° \times 4.0°$ for a movie, and $10.0° \times 2.0°$ for a search-icon. We determined these sizes by choosing a minimum target size and enlarging other targets appropriately to be able to understand their meaning. We

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

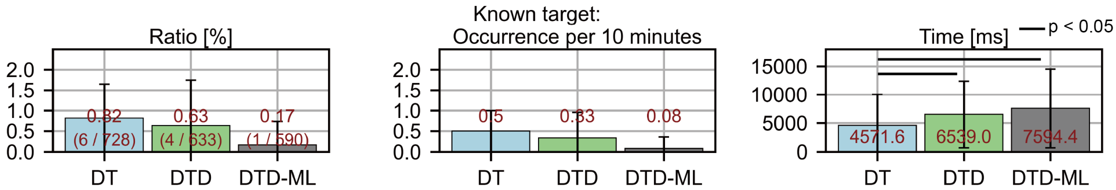

FIGURE 4.8: Quantitative results for the selection of known targets. Values in parentheses indicate numbers of unwanted selections and total selections.

chose the minimum size as 2.0°, which was approximately 2.3 cm on the screen used here, thereby making the size similar to that suggested in [FWT+17] (for filtered data, a target size of 1.9×2.35 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 2000 ms. The detection threshold was 0.800. We used the same ML model that gave the results for *all* (hyper-parameter) shown in Section 4.3.4.

4.4.3 Procedure

We asked each participant to calibrate the eye tracker before beginning the task. The order of the selection methods was randomized. We asked the participants to search for a target whose content was attractive and to select that target. We did not limit the method of searching and told them to interact freely with the interface. We asked the participants to interact for ten minutes for each selection method. 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 ten minutes of interaction, the participants answered the System Usability Scale (SUS) [Bro96] and the NASA-TLX [HS88] tests. They then rested for at least five minutes before moving to the next method. The experiment took an average of 53 min per participant. Each received JPY 5,000 (~USD 45).

4.4.4 Results

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 according to the number of total selections. The

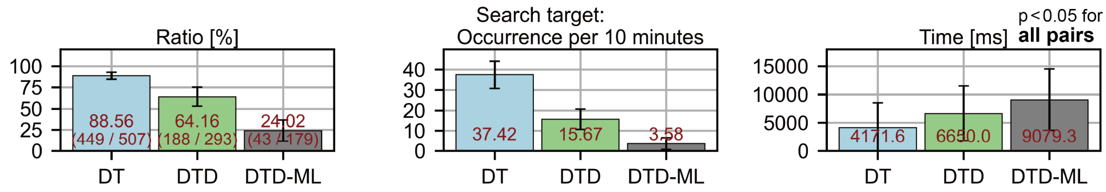

FIGURE 4.9: Quantitative results for the selection of search targets. Values in parentheses indicate number of unwanted selections and total selections.

time was calculated by subtracting the time at which the labeling form closed from the time at which a target was selected. Figures 4.8 and 4.9 show the results for selecting the known and search targets, respectively. For DT, DTD, and DTD-ML selections, the ratio and occurrence decreased in the order of DT, DTD, and DTD-ML selections, whereas the time increased in the 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 ratio, occurrence, and time. We found significant differences in the ratio and occurrence for search targets, which indicated that the screening of user intent with DTD detection and the intent detection with an ML model works well; DTD-ML selection (ratio: 24.02) prevented 40.2% of unwanted selection compared to DTD selection (ratio: 64.16), and DTD selection prevented 24.4% compared to DT selection (ratio: 88.56). For known targets, there were no significant differences in the ratio and occurrence. This confirms both the usefulness of DT selection for known targets and the result of the letter task in Experiment 1. In the case of time, there were significant differences between DT and the other selection methods for both known and search targets. Both the DTD and DTD-ML selection methods allowed the participants to search for a target more carefully. However, this also suggests that the DT selection allows faster selection compared to the DTD and DTD-ML selections.

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

Qualitative Results

Figures 4.10 and 4.11 show the NASA-TLX and SUS results, 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.

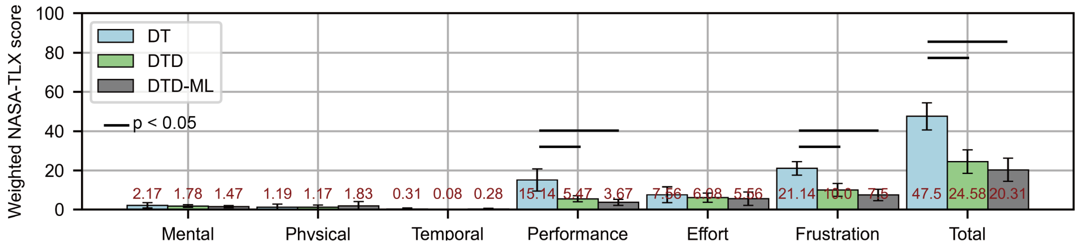

FIGURE 4.10: NASA-TLX test results; lower values indicate better scores.

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 the dwell interface did not require physical activity, the scores for the mental, physical, and temporal demands were smaller than the other scores. In terms of the performance and frustration scores, the 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 between the 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 increased in order. Regarding inconsistency, the DTD selection had the highest score. The 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 using each method, the scores became high with no significant differences. This indicates that the learning cost for dwell selection seems less regardless of the methods.

Detection Delays

We also measured the time required to create features and detect 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 detection were 3.55 ms [2.22–14.12]

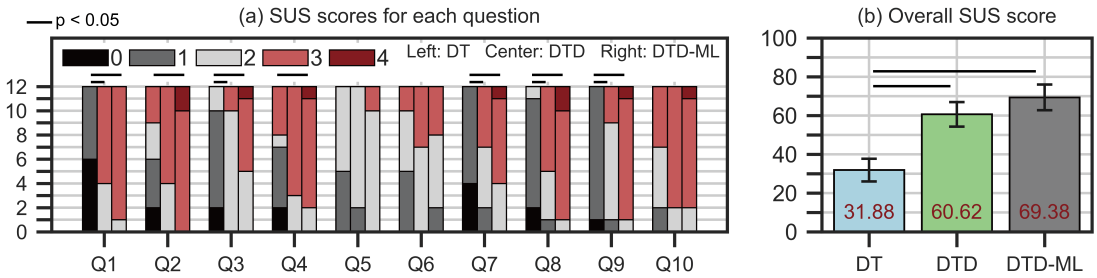

FIGURE 4.11: SUS test results. (a) Bar chart showing adjusted scores for each question in order of DT, DTD, and DTD-ML selection, where 0 (black) indicates the worst score and 4 (red) indicates the best score. (b) Box plot showing overall scores (higher is better) among participants.

and 0.21 ms [0.12–1.37], respectively. The delay in comparison to DTD selection averaged 3.76 ms [2.35–14.60]. As the eye-tracking frequency in Experiment 2 was 1200 Hz (i.e., 0.83 ms/sample), the detection could not be finished within one sample. However, when using DTD-ML selection for interaction, such a delay may not seem significant.

4.5 Discussion

4.5.1 Limitations on Applicable Interfaces and Interaction on DTD-ML

We showed that DTD-ML works for a dictionary-like interface whose contents are a size of at least 2.0° in size (2.3 cm in this experimental setting). We limited the target size to avoid issues related to eye-tracking accuracy. However, sizes smaller than 2.0° are used for tab-icons on the Windows 10 desktop and close buttons on a web browser. A target size of approximately 2.0° 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 a display zoom setting of 100%, and these contents are positioned in a grid layout with small margins between contents, which is similar to the interface used in Experiment 2. Therefore, the DTD-ML would work for a common interface design with a content size of at least 2.0°. Although the *leave-one-task-out* result possessed an insufficient AUC value, the results of Experiment 2, whose interface and task differed from those of Experiment 1, are robust against the Midas-touch problem. However, the capability for the selection of objects other than a character, word, sentence, or image is still unexplored; therefore, 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 a more realistic use of gaze-based interaction. One solution using DTD-ML would be the two-step manipulation, similar to right-clicking: the first selection would open a menu on a dwelled target, and the second selection would activate a command mapped to the dwelled menu item on the target. While this design is not new, since DTD-ML offers a robust trigger for opening a menu, it can prevent occlusion due to unwanted opening of the menu. This would also be useful for eye-gesture research, which uses dwell selection for trigger gesture detection (e.g., [ULH10, IYS20, DHI17, ASP+21b, KHAL22b]). Another solution would be to combine with a second modality (e.g., [PAC+15, PACG14, CXH15, PMMG17]). Note that the main contribution of this work is the establishment of an essential "selection" system like left-clicking a mouse, and hence, these limitations should be explored.

4.5.2 Participant Dependency

We achieved strong detection results for the participants considering 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 the 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, pupil diameter decreases with age [BCB50], and further investigation is needed to test our method on a diverse range of users.

4.5.3 Application to DT Selection

By changing the threshold of the detection 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 the research on tuning the dwell time to prevent the Midas-touch and achieve fast selection, and our work can contribute to this. 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, wherein only the dwell time is used. Thus, we can also apply our method to the research on DT selection (e.g., [ZRZ08, CSO20]) to improve performance.

4.5.4 Exploring Parameters

There is space for tuning the parameters used, e.g., the dwell time and dispersion threshold. The DTD detection roughly screens the user's intent to select; therefore, improvement in the accuracy of DTD detection would further alleviate the Midas-touch problem. The dwell time and dispersion threshold used were determined based on a preliminary investigation. Although we used 0.3° for the dispersion threshold, a lower value or the one adjusted for the target position would be ideal for improved screening.

As for the window size, as described in Section 4.3.4, a large window size yielded a high AUC; however, we should investigate the use of larger window sizes with data collected from a wider range of tasks. As described in Section 4.5.3, tuning the detection probability threshold would also improve the performance.

4.5.5 Feature Exploration

We reanalyze effective features for intent detection performance in different interaction situations. Owing to the myriad of interaction situations, it is challenging to examine them all exhaustively. As the first step toward understanding features for interactions, we retrained our model by changing features to explore how detection performance changes when different features are used. In particular, we focused on the perspective of dimensionality reduction and adaptation to different tasks.

This discussion is based on our work published on Eyes4ICU which is a workshop in ETRA 2023 [IYS23b].

Dimensionality Reduction We investigated features to improve our model in terms of dimensionality reduction. While many studies use saccade and fixation information as a primary feature indicating a user's intent, we found that the gaze movement and pupil changes were significant in detecting user intent in our model. Among the top 30 feature gains shown in Figure 4.6, only peak-to-peak saccade distance was considered as an important feature in saccade information. This indicates that saccade and fixation information may not be as important as other features for our model.

Therefore, we retrained an ML model with features that excluded saccade and fixation information, and we observed that the overall AUC improved from 0.903 to 0.904 while the number of features decreased from 127 to 92. While it is difficult to clearly explain the specific reasons behind the performance improvement due to the use of an ML approach, we speculate that the saccades behavior between the negative and positive classes did not differ significantly. For example, the distance of saccades was 4.5° and 3.9° in the negative and positive classes, respectively.

Here, 1.0° corresponds to 1.1 mm in the experimental environment, and 0.6° of difference may account for a small difference. Note that we used under 5.0° as the target size, which may account for both under 5.0° and a small difference (approx. 0.6°) in the distance. For an ML model detecting a user's intent that is unrelated to dwelling, such as [SZL+22], the saccade and fixation counts may vary even more significantly than those in our results. Therefore, despite many studies using saccade and fixation information as indicators of a user's intent, it is important to carefully consider the inclusion of such features.

Adaptation to Different Task Our model suffered from overfitting to the tasks, as demonstrated through the inadequate results of leave-one-task-out cross-validation (AUC=0.601). We examined the use of features in relation to the tasks. We utilized the gaze movement direction as a feature, represented by plus and minus values of $\mathbf{x}$ and $\mathbf{y}$. However, these values may be affected by various factors, including the type of content being viewed, the aspect ratio of the interface, size, and the arrangement of content. Therefore, the use of gaze movement direction concerning tasks must be carefully considered.

Consequently, we re-trained an ML model by excluding plus and minus values of changes in $\mathbf{x}$ and $\mathbf{y}$. Consequently, the AUC improved to 0.627 in the leave-one-task-out cross-validation, up from the initial AUC of 0.601; the overall AUC increased to 0.913 from 0.903. Although this improvement is still insufficient, excluding features that depend on the task can be a potential solution for overfitting that should be considered while developing an ML model.

4.6 Conclusion

We developed an ML-based model that detects user intent for selection with natural human eye behaviors. As features for the ML-based detection, we used gaze movement, fixation, saccade, pupil diameter, and vergence, which are linked to a user dwell action. To develop the intent detection model, we first conducted Experiment 1 on labeling user intent with five tasks and then calculated the features. The results showed that our model could detect a user's intent with a high AUC value of 0.903: specifically, 0.898 for detection independent of the user and 0.880 for detection independent of the eye tracker. The results of Experiment 2 showed that the DTD-ML selection could prevent 40.2% of unwanted selection compared to the DTD selection and that it yielded equal or better NASA-TLX and SUS scores than DT and DTD selection. Our approach to intent detection should significantly contribute to system development for various interactive situations, and further advancement based on our research may potentially allow the use of gaze-based intent detection.

Chapter 5

CONCLUSIONS

In this chapter, we discuss ways in which researchers can utilize our findings on gaze-based interaction and conclude this thesis.

5.1 Use of Our Dwell Time Determination Model

In the research on preventing Midas-touch, a faster and more accurate dwell selection has been developed (i.e., the best solution has been regarded as 0 ms of dwell time and zero Midas-touches); however, this seems to be difficult since no study has achieved this using dwell time-based user intent detection. However, if we can use a larger dwell time with a valid reason, there is a possibility that the solution is closer than now. For example, our model derives the dwell time, enabling dwell selection after a user completes the decision-making process required; we think such dwell time (i.e., 174 ms for a simple colored target selection task and 274 ms for other tasks) can be used as a target dwell time to achieve zero Midas-touches. Moreover, a dwell time smaller than the aforementioned dwell times potentially decreases the usability of dwell selection, as reported in previous studies [IAST18, CSO22]. Assuming that dwell time derived from our model does not decrease the usability of dwell selection, our model is helpful for future researchers addressing the Midas-touch.

Moreover, the dwell time that is determined based on our model can be used to determine the dwell time as one experimental condition. Researchers using dwell selection as a comparison method to evaluate the performance of interaction methods, such as dwell selection vs. explicit and multimodal gaze interaction (e.g., [CSO22, NAG$^+$23]) may utilize dwell time. The dwell times for method-comparison experiments were often determined through a preliminary study conducted in each research without detailed information. This is because there is no baseline dwell time that researchers can refer to, although dwell time is a parameter effect on a tradeoff between the speed and accuracy of dwell selection. We believe that researchers have not done this; however, if they want a result wherein their interaction system has a smaller error rate than dwell selection, they can adopt a small dwell time. As an example of dwell time as an experimental condition, we suggest using dwell times of at least 174.2, 350.4, 424.3, 651.9, and 835.8 ms for simple colored objects, key, icon, word, and image selection tasks, respectively. Of course, if the researchers consider the Midas-touch, a larger dwell time can be used; however, smaller dwell times are not appropriate unless experimental tasks do not require "searching" for a target.

In terms of extension for various interaction methods, our model may extend the implicit interaction, especially an interaction driven by user intent detection incorporating human natural behaviors. For example, there are studies on selection methods for GUI objects wherein the selection is done just before a user performs an explicit action of left-clicking [ASK+05, PW14, MW14]). Moreover, a recent study has shown that an interaction system automatically corrects an error input through intent detection using eye behaviors [PLZ+22]; the "undo" interaction that revokes the previously triggered interaction is triggered. Unfortunately, for these interactions, the time when such interactions should be triggered has not been investigated in detail. Similar to a recent study wherein it was reported that a shorter dwell time decreases usability, this time should also be carefully considered. In these scenarios, same as the dwell time in our work, we hope to adopt time determined by incorporating human eye behavior and the decision-making process.

5.2 Use of Our Intent Detection Model

We utilized the user intent detection model for dwell selection only. There have been various applications of eye behavior-based intent detection. For example, the area of interest is detected by using gaze coordinates, and the duration of the gaze stays at a point, similar to dwell time-based dwell detection. The area of interest is often used to create a heatmap of user interest to analyze the UI design. However, because dwell selection has failed when using a time threshold to detect user intent alone, the current detection of the area of interest successfully reflects whether or not the user's true interest is questionable. Moreover, we expect that our model will be useful for other interactions. The most promising application is the gaze-supported system combining gaze and other modalities (i.e., the multimodal gaze interaction), which other researchers have attempted to develop as the AR/VR interface (e.g., [DJPZ+21, PPE+21, LDB21]). 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 detect such varied intents owing to its current limitations in the types of detectable intent.

However, advancement based on this research should lead to further use of gaze-based intent detection and the development of real-world applications. We expect our user intent detection model or the methodology of developing the model to help detect more accurate user interest.

Dwell selection is often used as a condition for interaction method comparison experiments. However, because no dwell selection solves Midas-touch, all results on the performance of dwell selection are affected by Midas-touch. Consequently, researchers concluded that dwell selection has poor usability owing to the occurrence of Midas-touch or the necessity of looking more than necessary to prevent occurring Midas-touch. However, we think that this comparison is unfair from the aspect of dwell selection because they used dwell time even though they know Midas-touch occurs (actually, there has been no choice of Midas-touch free dwell selection). The experimental result may be changed if we use dwell selection where Midas-touch rarely occurs, such as our DTD-ML selection. Therefore, by using our intent detection model that prevents Midas-touch, we can evaluate interaction methods under more fair and ideal conditions again and can observe different findings, although the comparison has already been made in numerous research.

When utilizing ML-based intent detection that employs human eye behavior, such as the approach we have developed, for dwell selection with no Midas-touch, it becomes possible to set a dwell time of 0 ms as for dwell selection. This is because intent detection is based on the user's eye behavior prior to starting the dwell action. In this case, the size of dwell time roles the delay from when the user looks at an object to when the target is selected. In this context, the dwell time serves as the interval between the user's gaze entering an object and the subsequent selection of the target. Consequently, a dwell time of 0 ms indicates that the selection is triggered as soon as the system detects the user's gaze entering the target. However, as our 1st contribution involving the determination of dwell time based on the human decision-making process, it becomes apparent that a dwell time of 0 ms is not ideal. Introducing a certain delay has the potential to improve the usability of the dwell selection. For example, when selecting a simple colored object, we suggest that a minimum dwell time of 170 ms be selected for optimized results. Therefore, by incorporating our findings, the possibility arises to develop a dwell selection that mitigates Midas-touch while improving usability.

5.3 Conclusions

This thesis revealed how the user intent to either select or not select is detected using natural eye behaviors and established dwell selection as a daily interaction method.

In Chapter 3, we showed the development of a model that determined the dwell time from the relation between natural human eye behavior of fixation and the decision-making process, which is described in MHP. Because the decision-making process differs depending on the tasks, we conducted five selection tasks with different difficulties in completing the tasks to obtain eye behaviors during each selection task. Based on the analysis, we justified three hypotheses regarding the relation between fixation during the gaze-button selection task and MHP. The model results in fitting to the experimentally obtained data with over 0.9 of R^2 for all four tasks. Our model revealed dwell times for selecting an object that users fixate on it for the first time, for an object that users fixate on it at least two times; the smallest dwell time should be used to consider the human decision-making processes.

In Chapter 4, we showed the development of a model that detected user intent to select a target; the detection was based on ML that utilized features calculated from the natural eye behaviors during the dwell selection task. To develop the model, we conducted five tasks to obtain eye behavior from those tasks for the same reason as in the previous chapter. Based on the obtained eye behaviors, we developed the model. Our model could classify user intent to select or not to select with an AUC value of 0.903. The DTD-ML selection, which utilizes our intent detection model for dwell selection, prevented 40.2% and 90% of Midas-touches compared to DTD and DT selections, respectively. Moreover, we demonstrated that the DTD selection yielded equal or better qualitative results of NASA-TLX and SUS scores than the DT and DTD selections.

Lastly, we demonstrated how the two models can be used for future interaction on gaze-based interaction.

This thesis reports two models for the detection of user intent to interact with a computer. However, we have only scratched the surface of how natural human eye behavior can be used to reveal user intent for gaze-based interaction. There are more functions and characteristics in natural human eye behaviors that were not considered in this thesis. Considering our work is a first step toward an in-depth understanding of the implicitness of natural eye behaviors for gaze-based interaction, we believe our work has opened a new pathway that extends toward becoming gaze-based interaction as a common interaction method.

Bibliography

[Aka74] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[AL13] Borji Ali and Itti Laurent. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:185–207, 2013.

[AMD95] John R. Anderson, Michael Matessa, and Scott A. Douglass. The act-r theory and visual attention. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, pages 61–65. Hillsdale, NJ: Lawrence Erlbaum Associates, 1995.

[AML97] John R. Anderson, Michael Matessa, and Christian Lebiere. ACT-R: A Theory of Higher Level Cognition and its Relation to Visual Atention. *Human-Computer Interaction*, 12(4):439–462, 1997.

[ASK+05] Takeshi Asano, Ehud Sharlin, Yoshifumi Kitamura, Kazuki Takashima, and Fumio Kishino. Predictive Interaction Using the Delphian Desktop. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, UIST '05, pages 133–141. Association for Computing Machinery, 2005.

[ASLL20] Sunggeun Ahn, Jeongmin Son, Sangyoon Lee, and Geehyuk Lee. 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*, CHI EA '20, pages 264:1–7. Association for Computing Machinery, 2020.

[ASP+21a] Sunggeun Ahn, Stephanie Santosa, Mark Parent, Daniel Wigdor, Tovi Grossman, and Marcello Giordano. 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*, New York, NY, USA, 2021. Association for Computing Machinery.

[ASP+21b] Sunggeun Ahn, Stephanie Santosa, Mark Parent, Daniel Wigdor, Tovi Grossman, and Marcello Giordano. 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, 2021.

[ASY+19] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631. Association for Computing Machinery, 2019.

[BA03] Kenneth P Burnham and David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer Science & Business Media, Heidelberg, Germany, 2003.

[BCB50] James E. Birren, Roland C. Casperson, and Jack Botwinick. Age Changes in Pupil Size. *Journal of Gerontology*, 5(3):216–221, July 1950.

[BOBH14] Gilles Bailly, Antti Oulasvirta, Duncan P. Brumby, and Andrew Howes. Model of Visual Search and Selection Time in Linear Menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3865–3874. Association for Computing Machinery, 2014.

[Bra97] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[Bro96] John Brooke. *Usability Evaluation in Industry*, chapter SUS-A Quick and Dirty Usability Scale, pages 189–194. CRC Press, 1996.

[BVH12] Roman Bednarik, Hana Vrzakova, and Michal Hradis. 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*, ETRA '12, pages 83–90. Association for Computing Machinery, 2012.

[CEU03] P.P. Caffier, U. Erdmann, and P. Ullsperger. Experimental evaluation of eye-blink parameters as a drowsiness measure. *European Journal of Applied Physiology*, 89:319–325, 2003.

[CGG07] Andy Cockburn, Carl Gutwin, and Saul Greenberg. A Predictive Model of Menu Performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 627–636. Association for Computing Machinery, 2007.

[Cly62] Manfred Clynes. 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):806–845, 1962.

[CMN80] Stuart K. Card, Thomas P. Moran, and Allen Newell. The Keystroke-Level Model for User Performance Time with Interactive Systems. *Communications of the ACM*, 23(7):396–410, July 1980.

[CNM83] Stuart K. Card, Allen Newell, and Thomas P. Moran. *The Psychology of Human-Computer Interaction*, chapter 2. L. Erlbaum Associates Inc., USA, 1983.

[CSO20] Myungguen Choi, Daisuke Sakamoto, and Tetsuo Ono. Bubble Gaze Cursor + Bubble Gaze Lens: Applying Area Cursor Technique to Eye-Gaze Interface. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Full Papers, pages 1–10, New York, NY, USA, 2020. Association for Computing Machinery.

[CSO22] Myungguen Choi, Daisuke Sakamoto, and Tetsuo Ono. Kuiper Belt: Utilizing the "Out-of-Natural Angle" Region in the Eye-Gaze Interaction for Virtual Reality. In *CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, 2022.

[CXH15] Ishan Chatterjee, Robert Xiao, and Chris Harrison. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 131–138. Association for Computing Machinery, 2015.

[DHI17] William Delamare, Teng Han, and Pourang Irani. Designing a Gaze Gesture Guiding System. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, pages 26:1–26:13. Association for Computing Machinery, 2017.

[DJPZ+21] Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. Towards Gaze-Based Prediction of the Intent to Interact in Virtual Reality. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '21 Short Papers. Association for Computing Machinery, 2021.

[DKA18] Heiko Drewes, Mohamed Khamis, and Florian Alt. Smooth Pursuit Target Speeds and Trajectories. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, MUM '18, pages 139–146. Association for Computing Machinery, 2018.

[DS07] Heiko Drewes and Albrecht Schmidt. Interacting with the Computer Using Gaze Gestures. In *Proceedings of the 11th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part II*, INTERACT '07, pages 475–488, Berlin, Heidelberg, 2007. Springer-Verlag.

[Dyn21] Tobii Dynavox. Assistive technology for communication/AAC - Tobii Dynavox, 2021. (Retrieved January 27, 2021).

[EVBG15] Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. 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*, UIST '15, pages 457–466. Association for Computing Machinery, 2015.

[FCHN15] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970. IEEE, 2015.

[FF18] Pedro Figueiredo and Manuel J. Fonseca. EyeLinks: A Gaze-Only Click Alternative for Heterogeneous Clickables. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 307â 314. Association for Computing Machinery, 2018.

[Fit54] P. M. Fitts. The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. *Journal of Experimental Psychology*, 74:381–391, 1954.

[FWT+17] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 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*, CHI '17, pages 1118–1130. Association for Computing Machinery, 2017.

[GBL+03] K. Grauman, M. Betke, J. Lombardi, J. Gips, and G. R. Bradski. Communication via Eye Blinks and Eyebrow Raises: Video-Based Human-Computer Interfaces. *Universal Access in the Information Society*, 2(4):359â 373, November 2003.

[HB05] Mary Hayhoe and Dana Ballard. Eye Movements in Natural Behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005.

[HC05] Anthony J. Hornof and Anna Cavender. EyeDraw: Enabling Children with Severe Motor Impairments to Draw with Their Eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 161–170. Association for Computing Machinery, 2005.

[HCH04] Anthony Hornof, Anna Cavender, and Rob Hoselton. EyeDraw: A System for Drawing Pictures with the Eyes. In *Proceedings of the 2004 CHI Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, pages 1251–1254. Associati0on for Computing Machinery, 2004.

[HJH+03] John Hansen, Anders Johansen, Dan Hansen, Kenji Ito, and Satoru Mashino. Command Without a Click: Dwell Time Typing by Mouse and Gaze Selections. In *Proceedings of Human-Computer Interaction*, INTERACTA '03, pages 121–128. International Federation for Information Processing, 2003.

[HMAR00] Aulikki Hyrskykari, Päivi Majaranta, Antti Aaltonen, and Kari-Jouko Räihä. Design Issues of IDICT: A Gaze-Assisted Translation Aid. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pages 9–14, New York, NY, USA, 2000. Association for Computing Machinery.

[HP60] Eckhard H. Hess and James M. Polt. Pupil Size as Related to Interest Value of Visual Stimuli. *Science*, 132:349–350, 1960.

[HS88] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[HU08] Anke Huckauf and Mario H. Urbina. Gazing with PEYEs: Towards a Universal Input for Various Applications. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, ETRA '08, pages 51–54. Association for Computing Machinery, 2008.

[HWM+89] Thomas E. Hutchinson, K. Preston White, Worthy N. Martin, Kelly C. Reichert, and Lisa A. Frey. Human-computer Interaction using Eye-gaze Input. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1527–1534, 1989.

[IAST18] Toshiya Isomoto, Toshiyuki Ando, Buntarou Shizuki, and Shin Takahashi. Dwell Time Reduction Technique Using Fitts' Law for Gaze-Based Target Acquisition. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA '18, pages 26:1–26:7. Association for Computing Machinery, 2018.

[IHI+10] Howell Istance, Aulikki Hyrskykari, Lauri Immonen, Santtu Mansikkamaa, and Stephen Vickers. Designing Gaze Gestures for Gaming: An Investigation of Performance. In *Proceedings of the 2010 ACM Symposium on Eye-Tracking Research & Applications*, ETRA '10, pages 323–330. Association for Computing Machinery, 2010.

[IMKD20] Shoya Ishimaru, Takanori Maruichi, Koichi Kise, and Andreas Dengel. Gaze-Based Self-Confidence Estimation on Multiple-Choice Questions and Its Feedback. In *Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures*, AsianCHI '20, page 8, New York, NY, USA, 2020. Association for Computing Machinery.

[IYS20] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. Gaze-based Command Activation Technique Robust Against Unintentional Activation using Dwell-then-Gesture. In *Proceedings of Graphics Interface 2020*, GI '20, pages 256–266. Canadian Human-Computer Communications Society / SociÃ©tÃ© canadienne du dialogue humain-machine, 2020.

[IYS21] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. Relationship between Dwell-Time and Model Human Processor for Dwell-based Image Selection. In *Proceedings of the 2021 ACM Symposium on Applied Perception*, SAP '21, pages 1–5. Association for Computing Machinery, 2021.

[IYS22] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. Dwell Selection with ML-Based Intent Prediction Using Only Gaze Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–21, September 2022.

[IYS23a] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. Exploring Dwell-Time from Human Cognitive Processes for Dwell Selection. *Proceedings of the ACM on Human-Computer Interaction*, 7(ETRA):1–15, May 2023.

[IYS23b] Toshiya Isomoto, Shota Yamanaka, and Buntarou Shizuki. Reanalyzing Effective Eye-Related Information for Developing User’ s Intent Detection Systems. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, ETRA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[Jac90] Robert J. K. Jacob. 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*, CHI '90, pages 11–18. Association for Computing Machinery, 1990.

[JHF17] Florian Jungwirth, Michael Haslgrübler, and Alois Ferscha. Contour-guided gaze gestures: Eye-based interaction with everyday objects and iot devices. In *Proceedings of the Seventh International Conference on the Internet of Things*, IoT '17, pages 26:1–26:2. Association for Computing Machinery, 2017.

[JSSV15] Joaquin Jadue, Gino Slanzi, Luis Salas, and Juan D. VelÃ¡squez. Web user click intention prediction by using pupil dilation analysis. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1 of *WI-IAT '15*, pages 433–436, New York, NY, USA, 2015. IEEE / ACM.

[JW15] Ricardo Jota and Daniel Wigdor. Palpebrae Superioris: Exploring the Design Space of Eyelid Gestures. In *Proceedings of the 41st Graphics Interface Conference*, GI '15, pages 273–280, CAN, 2015. Canadian Information Processing Society.

[KAH+16] Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. 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*, CHI EA '16, pages 2156–2164. Association for Computing Machinery, 2016.

[KB16] Dominik Kirst and Andreas Bulling. 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*, CHI EA '16, pages 1519–1525. Association for Computing Machinery, 2016.

[KHAL22a] Taejun Kim, Auejin Ham, Sunggeun Ahn, and Geehyuk Lee. Lattice Menu: A Low-Error Gaze-Based Marking Menu Utilizing Target-Assisted Gaze Gestures on a Lattice of Visual Anchors. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[KHAL22b] Taejun Kim, Auejin Ham, Sunggeun Ahn, and Geehyuk Lee. 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*, CHI '22. Association for Computing Machinery, 2022.

[KMS10] Dagmar Kern, Paul Marshall, and Albrecht Schmidt. Gazemarks: Gaze-Based Visual Placeholders to Ease Attention Switching. In *Proceedings of the 2010 CHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2093–2102. Association for Computing Machinery, 2010.

[KNBV22] Anam Ahmad Khan, Joshua Newn, James Bailey, and Eduardo Velloso. Integrating gaze and speech for enabling implicit interactions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[KOH+13] Shinya Kudo, Hiroyuki Okabe, Taku Hachisu, Michi Sato, Shogo Fukushima, and Hiroyuki Kajimoto. Input Method Using Divergence Eye Movement. In *Proceedings of the 2013 CHI Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 1335–1340. Association for Computing Machinery, 2013.

[KS19] Piotr Kowalczyk and Dariusz Sawicki. Blink and Wink Detection as a Control Tool in Multimodal Interaction. *Multimedia Tools and Applications*, 78(10):13749â 13765, May 2019.

[Kur93] Gordon Kurtenbach. *The Design and Evaluation of Marking Menus.* PhD thesis, Toronto, Ont., Canada, Canada, 1993. UMI Order No. GAXNN-82896.

[LDB21] Feiyu Lu, Shakiba Davari, and Doug Bowman. Exploration of Techniques for Rapid Activation of Glanceable Information in Head-Worn Augmented Reality. In *Symposium on Spatial User Interaction*, SUI '21, pages 1–14, New York, NY, USA, 2021. Association for Computing Machinery.

[LH01] Michael F. Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25):3559–3565, 2001.

[LPW15] Christof Lutteroth, Moiz Penkar, and Gerald Weber. Gaze vs. Mouse: A Fast and Accurate Gaze-Only Click Alternative. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, UIST '15, pages 385–394. Association for Computing Machinery, 2015.

[Mac91] Ian Scott Mackenzie. *Fitts' Law As a Performance Model in Human-computer Interaction.* PhD thesis, Toronto, Ont., Canada, Canada, 1991.

[MAR04] Päivi Majaranta, Anne Aula, and Kari-Jouko Räihä. Effects of Feedback on Eye Typing with a Short Dwell Time. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, ETRA '04, pages 139–146. Association for Computing Machinery, 2004.

[MAv09] Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. Fast Gaze Typing with an Adjustable Dwell Time. In *Proceedings of the 2009 CHI Conference on Human Factors in Computing Systems*, CHI '09, pages 357–360. Association for Computing Machinery, 2009.

[MB10] Eric Missimer and Margrit Betke. Blink and Wink Detection for Mouse Pointer Control. In *Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '10, New York, NY, USA, 2010. Association for Computing Machinery.

[MGFY18] Pallavi Mohan, Wooi Boon Goh, Chi-Wing Fu, and Sai-Kit Yeung. 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, pages 79–84, 2018.

[MHL13a] Emilie MÃ¸llenbach, John Paulin Hansen, and Martin Lillholm. Eye movements in gaze interaction. *Journal of Eye Movement Research*, 6(2), May 2013.

[MHL13b] Emilie Møllenbach, John Paulin Hansen, and Martin Lillholm. Eye Movements in Gaze Interaction. *Journal of Eye Movement Research*, 6(2):1–15, 2013.

[MHLG09] Emilie Møllenbach, John Paulin Hansen, Martin Lillholm, and Alastair G. Gale. Single Stroke Gaze Gestures. In *Proceedings of the 2009 CHI Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, pages 4555–4560. Association for Computing Machinery, 2009.

[MLGH10] Emilie Møllenbach, Martin Lillholm, Alastair Gail, and John Paulin Hansen. Single Gaze Gestures. In *Proceedings of the 2010 ACM Symposium on Eye-Tracking Research & Applications*, ETRA '10, pages 177–180. Association for Computing Machinery, 2010.

[MLH20] Sven Mayer, Gierad Laput, and Chris Harrison. Enhancing Mobile Voice Assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–10. Association for Computing Machinery, 2020.

[MMAR06] Päivi Majaranta, I. Scott MacKenzie, Anne Aula, and Kari-Jouko Räihä. Effects of Feedback and Dwell Time on Eye Typing Speed and Accuracy. *Universal Access in the Information Society*, 5(2):199–208, 2006.

[MW14] Martez E. Mott and Jacob O. Wobbrock. Beating the Bubble: Using Kinematic Triggering in the Bubble Lens for Acquiring Small, Dense Targets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 733–742. Association for Computing Machinery, 2014.

[MWWM17] Martez E. Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2558–2570. Association for Computing Machinery, 2017.

[NAG+23] Omar Namnakani, Yasmeen Abdrabou, Jonathan Grizou, Augusto Esteves, and Mohamed Khamis. Comparing Dwell Time, Pursuits and Gaze Gestures for Gaze Interaction on Handheld Mobile Devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[NDA+17] Aanand Nayyar, Utkarsh Dwivedi, Karan Ahuja, Nitendra Rajput, Seema Nagar, and Kuntal Dey. OptiDwell: Intelligent Adjustment of Dwell Click Time. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 193–204. Association for Computing Machinery, 2017.

[NSA+23] Omar Namnakani, Penpicha Sinrattanavong, Yasmeen Abdrabou, Andreas Bulling, Florian Alt, and Mohamed Khamis. GazeCast: Using Mobile Devices to Allow Gaze-Based Interaction on Public Displays. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, ETRA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[PAC+15] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, Yanxia Zhang, and Hans Gellersen. 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*, UIST '15, pages 373–383. Association for Computing Machinery, 2015.

[PACG14] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, and Hans Gellersen. 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*, UIST '14, pages 509–518. Association for Computing Machinery, 2014.

[PL18] Ken Pfeuffer and Yang Li. Analysis and Modeling of Grid Performance on Touchscreen Mobile Devices. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–12. Association for Computing Machinery, 2018.

[PLLB17] Thammathip Piumsomboon, Gun Lee, Robert W. Lindeman, and Mark Billinghurst. Exploring Natural Eye-Gaze-based Interaction for Immersive Virtual Reality. In *2017 IEEE Symposium on 3D User Interfaces*, 3DUI '17, pages 36–39, 2017.

[PLW13] Abdul Moiz Penkar, Christof Lutteroth, and Gerald Weber. Eyes Only: Navigating Hypertext with Gaze. In *14th IFIP TC 13 International Conference on Human-Computer Interaction – INTERACT 2013*, pages 153–169, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[PLZ+22] Candace E. Peacock, Ben Lafreniere, Ting Zhang, Stephanie Santosa, Hrvoje Benko, and Tanya R. Jonker. Gaze as an Indicator of Input Recognition Errors. *Proceedings of ACM Human-Computer Interaction*, 6(ETRA), May 2022.

[PMMG17] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. Gaze + Pinch Interaction in Virtual Reality. In *Proceedings of the 5th Symposium on Spatial User Interaction*, SUI '17, pages 99–108. Association for Computing Machinery, 2017.

[PPE+21] Robin Piening, Ken Pfeuffer, Augusto Esteves, Tim Mittermeier, Sarah Prange, Philippe Schröder, and Florian Alt. 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*, pages 544–565. Springer International Publishing, 2021.

[PS17] Jimin. Pi and Bertram. E. Shi. Probabilistic Adjustment of Dwell Time for Eye Typing. In *10th International Conference on Human System Interactions (HSI)*, pages 251–257. IEEE, 2017.

[PSD12] Panwar Prateek, Sarcar Sayan, and Samanta Debasis. EyeBoard: A Fast and Accurate Eye Gaze-Based Text Entry System. In *2012 4th*

International Conference on Intelligent Human Computer Interaction, IHCI '12, pages 1–8, 2012.

[PW14] Phillip T. Pasqual and Jacob O. Wobbrock. Mouse Pointing Endpoint Prediction Using Kinematic Template Matching. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 743–752. Association for Computing Machinery, 2014.

[RGCSG21] Argenis Ramirez Ramirez Gomez, Christopher Clarke, Ludwig Sidenmark, and Hans Gellersen. Gaze+Hold: Eyes-Only Direct Manipulation with Continuous Gaze Modulated by Closure of One Eye. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '21 Full Papers, New York, NY, USA, 2021. Association for Computing Machinery.

[RH18] Vijay Rajanna and Tracy Hammond. A gaze gesture-based paradigm for situational impairments, accessibility, and rich interactions. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA '18, pages 102:1–102:3. Association for Computing Machinery, 2018.

[RO12] Kari-Jouko Räihä and Saila Ovaska. 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*, CHI '12, pages 3001–3010. Association for Computing Machinery, 2012.

[SA00] Dario D. Salvucci and John R. Anderson. Intelligent Gaze-Added Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 273–280. Association for Computing Machinery, 2000.

[SCN+23] Ludwig Sidenmark, Christopher Clarke, Joshua Newn, Mathias N. Lystbæk, Ken Pfeuffer, and Hans Gellersen. Vergence Matching: Inferring Attention to Objects in 3D Environments for Gaze-Assisted Selection. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[SD12a] Sophie Stellmach and Raimund Dachselt. Investigating Gaze-Supported Multimodal Pan and Zoom. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, pages

357–360, New York, NY, USA, 2012. Association for Computing Machinery.

[SD12b] Sophie Stellmach and Raimund Dachselt. Look & Touch: Gaze-supported Target Acquisition. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2981–2990. Association for Computing Machinery, 2012.

[SD13] Sophie Stellmach and Raimund Dachselt. Still Looking: Investigating Seamless Gaze-Supported Selection, Positioning, and Manipulation of Distant Targets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 285–294, New York, NY, USA, 2013. Association for Computing Machinery.

[SDRD17] Simon Schenk, Marc Dreiser, Gerhard Rigoll, and Michael Dorr. 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*, CHI '17, pages 3034–3044. Association for Computing Machinery, 2017.

[SG00] Dario D. Salvucci and Joseph H. Goldberg. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pages 71–78. Association for Computing Machinery, 2000.

[SG19a] Ludwig Sidenmark and Hans Gellersen. Eye, Head and Torso Coordination During Gaze Shifts in Virtual Reality. *ACM Transactions on Computer-Human Interaction*, 27(1), December 2019.

[SG19b] Ludwig Sidenmark and Hans Gellersen. Eye&Head: Synergetic Eye and Head Movement for Gaze Pointing and Selection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, pages 1161–1174. Association for Computing Machinery, 2019.

[SGBG08] Robert Schleicher, Niels Galley, Susan G. Briest, and Lars Arnim Galley. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, 51:982–1010, 2008.

[ŠIK+16] Oleg Špakov, Poika Isokoski, Jari Kangas, Deepak Akkil, and Päivi Majaranta. PursuitAdjuster: An Exploration into the Design Space of Smooth Pursuit-based Widgets. In *Proceedings of the 2016 ACM*

Symposium on Eye Tracking Research & Applications, ETRA '16, pages 287–290. Association for Computing Machinery, 2016.

[SJ00] Linda E. Sibert and Robert J. K. Jacob. Evaluation of Eye Gaze Interaction. In *Proceedings of the 2000 CHI Conference on Human Factors in Computing Systems*, CHI '00, pages 281–288. Association for Computing Machinery, 2000.

[SLW19] Asma Shakil, Christof Lutteroth, and Gerald Weber. CodeGazer: Making Code Navigation Easy and Natural With Gaze Input. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12. Association for Computing Machinery, 2019.

[SRT11] Henrik Skovsgaard, Kari-Jouko Räihä, and Martin Tall. Computer Control by Gaze. In Päivi Majaranta, Hirotaka Aoki, Mick Donegan, Witzner Hansen Dan, John Paulin Hansen, Aulikki Hyrskykari, and Kari-Jouko Räihä, editors, *Gaze Interaction and Aplications of Eye Tracking: Advances in Assistive Technologies*, chapter 9, pages 78–103. IGI Global, Hershey, PA, 2011.

[SZL+22] Naveen Sendhilnathan, Ting Zhang, Ben Lafreniere, Tovi Grossman, and Tanya R. Jonker. Detecting Input Recognition Errors and User Errors Using Gaze Dynamics in Virtual Reality. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery.

[TA08] Geoffrey Tien and M. Stella Atkins. Improving Hands-Free Menu Selection Using Eyegaze Glances and Fixations. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, ETRA '08, pages 47–50. Association for Computing Machinery, 2008.

[TABG15] Jayson Turner, Jason Alexander, Andreas Bulling, and Hans Gellersen. 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*, CHI '15, pages 4179–4188. Association for Computing Machinery, 2015.

[ULH10] Mario H. Urbina, Maike Lorenz, and Anke Huckauf. Pies with EYEs: The Limits of Hierarchical Pie Menus in Gaze Control. In *Proceedings* *of the 2010 ACM Symposium on Eye-Tracking Research & Applications*, ETRA '10, pages 93–96. Association for Computing Machinery, 2010.

[VBG13] Mélodie Vidal, Andreas Bulling, and Hans Gellersen. 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*, UbiComp '13, pages 439–448. Association for Computing Machinery, 2013.

[VCKM18] Eduardo Velloso, Flavio Luiz Coutinho, Andrew Kurauchi, and Carlos H Morimoto. 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*, ETRA '18, pages 25:1–25:9. Association for Computing Machinery, 2018.

[Ver03] Roel Vertegaal. Attentive User Interfaces. *Communications of the ACM*, 46(3), March 2003.

[vM04] Oleg Špakov and Darius Miniotas. On-Line Adjustment of Dwell Time for Target Selection by Gaze. In *Proceedings of the 2004 Nordic Conference on Human-Computer Interaction*, NordiCHI '04, pages 203–206. Association for Computing Machinery, 2004.

[Wei91] Mark Weiser. The Computer for the 21st Century. *Scientific American*, 265(3):94–105, 1991.

[Wid84] Heino Widdel. Operational Problems in Analysing Eye Movements. In Alastair G. Gale and Frank Johnson, editors, *Theoretical and Applied Aspects of Eye Movement Research*, volume 22 of *Advances in Psychology*, pages 21–29. North-Holland, 1984.

[WM87] Colin Ware and Harutune H. Mikaelian. 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*, CHI '87, pages 183–188. Association for Computing Machinery, 1987.

[WRSD08] Jacob O. Wobbrock, James Rubinstein, Michael W. Sawyer, and Andrew T. Duchowski. Longitudinal Evaluation of Discrete Consecutive

Gaze Gestures for Text Entry. In *Proceedings of the 2008 ACM Symposium on Eye Tracking Research & Applications*, ETRA '08, pages 11–18. Association for Computing Machinery, 2008.

[XSB16] Pingmei Xu, Yusuke Sugano, and Andreas Bulling. *Spatio-Temporal Modeling and Prediction of Visual Attention in Graphical User Interfaces*, pages 3299–3310. Association for Computing Machinery, 2016.

[ZIGM04] Qiaohui Zhang, Atsumi Imamiya, Kentaro Go, and Xiaoyang Mao. Resolving ambiguities of a gaze and speech interface. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, ETRA '04, page 85â 92, New York, NY, USA, 2004. Association for Computing Machinery.

[ZMI99] Shumin Zhai, Carlos Morimoto, and Steven Ihde. Manual and Gaze Input Cascaded (MAGIC) Pointing. In *Proceedings of the 1999 CHI Conference on Human Factors in Computing Systems*, CHI '99, pages 246–253. Association for Computing Machinery, 1999.

[ZRZ08] Xinyong Zhang, Xiangshi Ren, and Hongbin Zha. Improving eye cursor's stability for eye pointing tasks. In *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems*, CHI '08, pages 525–534. Association for Computing Machinery, 2008.

[ZXZZ11] Xinyong Zhang, Pianpian Xu, Qing Zhang, and Hongbin Zha. 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*, PETMEI '11, pages 37–42. Association for Computing Machinery, 2011.

APPENDIX

Supplementary for Chapter 3

TABLE 5.1: 220 Words used in word task.

Accessibility	Format Painter	Recording
Add Cursor	Formatting	Rectangle
Add Image	From CSV	Reference
Add Item	From Range	Related Dates
Add List	From Table	Release Note
Add Text	From Text	Remove Arrows
Applications	From Web	Remove Duplicates
Arrange All	Full Screen	Report Issue
Asian Layout	Get Help	Revert File
Auto Save	Get Started	Save As
AutoFormat	Gridlines	Save Document
AutoText	Help Center	Save File
Background	Highlighting	Save Image
Bluetooth	Histogram	Save Text
Bold Italic	Hyperlink	Searching
Bookmark	Increase Indent	Select All
Bring Forward	Insert Cell	Selection Pane
Calculate Now	Insert Function	Send Backward
Calculate Sheet	Join Channel	Sheet Options
Calculation Options	Join Room	Shortcut
Change Case	Last Modified	Show Comment
Change Icon	Last Printed	Show Comments
Change Name	Membership	Show Formulas
Character Count	Merge Cells	Show Icon
Check Accessibility	More Functions	Show Image
Clipboard	Mouse Pointer	Show Minimap
Close Editor	Move Line	Show Text

TABLE 5.1: (continued)

Close File	My Account	Shrink Selection
Close Folder	Name Manager	Shut Down
Close Text	Negative Numbers	Spell Check
Close Window	New Comment	Split Cells
Collaboration	New Folder	Split Down
Component	New Page	Split Left
Connection	New Terminal	Split Right
Contact Us	New Text	Split Table
Copy Line	New Window	Split Terminal
Create Shortcut	Next Comment	Spreadsheet
Custom Footer	Next Editor	Start-Up
Custom Header	Next Page	Strikethrough
Customize	Notification	Summarize
Data Types	Open File	Superscript
Date Time	Open Folder	Switch To
Decimal Places	Open Here	Task Manager
Decrease Indent	Open Recent	Task Pane
Define Name	Open Text	Text Alignment
Delete All	Organization	Text Box
Delete Cell	Overwrite	Text Color
Delete File	Page Numbers	Text Cursor
Delete Item	Page Setup	Text Direction
Dictionary	Page Setup	Text Size
Disable All	Paragraph	Three Columns
Document	Paste Special	Three Rows
Document Map	Permission	Thumbnails
Don't Show	Personalization	Toolbars
Download	Photography	ToolBox
Draw Table	Plot Graph	Track Change
Duplicate Directory	Prev Page	Translate
Duplicate File	Previous Comment	Two Columns
Duplicate Item	Previous Editor	Two Rows
Duplicate Letter	Print Area	Type Here
Duplicate Line	Print Layout	Underline
Duplicate Text	Print Preview	Uninstall
Duplicate Word	Print Titles	Version History
Enable All	Privacy Statement	Video Tutorial
Environment	Project setting	View License

Table 5.1: (continued)

Error Checking	Proofread	Watch Window
Evaluate Formula	Properties	Web Browser
Expand Selection	Preference	Web Capture
File Search	Quick Access	Web Layout
Fit Text	Recent File	Web Page
Flash Fill	Recent Source	Word Count
Flip Layout	Recently Used	Word Wrap
Footnotes	Recommendation	Workspace
Foreground		

FIGURE 5.1: Icons and instructions used for icon task.

Table 5.2: Three-layer hierarchical menu used for the word task.

First layer	Second layer	Third layer
Country	Asia	Japan, Korea, China, Thailand
	Europe	France, England, Germany, Spain
	America	Canada, United States, Cuba, Mexico
	Africa	Egypt, Ghana, Ethiopia, Kenya
Animal	Fish	Salmon, Lobster, Tuna, Octopus
	Insect	Ant, Bee, Ladybug, Beetle
	Mammal	Gorilla, Monkey, Dog, Horse
	Bird	Duck, Crow, Sparrow, Hawk
Drink	Alcohol	Wine, Beer, Whiskey, Sake
	Non-Alcohol	Water, Cocoa, Milk, Coffee
	Fruit	Orange, Apple, Peach, Grape
	Tea	Earl Grey, Darjeeling, Green Tea, Assam
Edit	File	Copy, Paste, Open, Undo
	Color	Red, Green, Blue, White
	Window	Show, Close, Show All, Close All
	Option	Preference, Account, Language, Help
Month	Spring	March, April, May
	Summer	June, July, August
	Fall	September, October November
	Winter	December, January, February

Supplementary for Chapter 4

Table 5.3: 300 sentences used in sentence task in Experiment 1.

Instruction	Sentence
一寸先は闇	将来のことは、ほんのわずか先のことですら、全くわからないということ。
生殺与奪	生かすも殺すも、与えるも奪うも、どの様にしようと思うがままであること
病は気から	病気は、本人の気持ちの持ち方次第で、重くもなるし軽くもなるということ。
泣きっ面に蜂	悪い目にあっているとき更に別の悪い目にあうこと。不幸や災難が重なること。

Table 5.3: (continued)

Instruction	Sentence
手前味噌	自分で自分を褒めること。自己の行動について卑下の表現として用いる例が多い
能ある鷹は爪を隠す	優れた能力のある人はそれを無駄にひけらかしたりしないということのたとえ。
残り物には福がある	人が取り残したものや、最後に残ったものの中には、思いがけず良いものがある。
魑魅魍魎	得体の知れない怪物、妖怪。また、それに類するもの。魑魅も魍魎も化け物の意。
頼みの綱	頼りにしてすがる物や人。もはやそれ以外にすがるものがない時に言うことが多い
百発百中	矢や銃弾が、みな的にあたること。予想やねらいなどがすべて思いどおりになること
一難去ってまた一難	一つの災難が過ぎてすぐに別の災難が降りかかること。次々に災難が襲ってくること。
馬子にも衣装	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。
朝三暮四	本質は変わらないのに、口先でうまくだます、又は、だまされることその愚かさのたとえ
虎視眈々	虎が目を見張って、獲物を狙う様、転じて、実力ある者が、じっと機会を伺っている様子。
憎まれっ子世にはばかる	他人に嫌われるくらいの人の方が、世に出た後に、幅をきかせることができるものである。
背水の陣	これ以上下がれない状態で、必死に物事を行うこと。後がない状態に身を置く、置かれること。
為せば成る	思案ばかりして、成果をあげようとする行動を起こさなければ、決して成果を得ることはない。
千差万別	種別がとても多いこと。種々様々で、実に様々な違いがあること。またはその様子。千種万様。

TABLE 5.3: (continued)

Instruction	Sentence
海老で鯛を釣る	（高価なタイを安いエビで釣るところから）少量の元手やわずかな労力で大きな利益を得ること。
暗中模索	先が見えず、決まった方向性が無い状況で、様々な行動に取り組んで事態を打開しようとすること
他力本願	他人任せで自分の望みを叶えようとする事。自分で努力をしないことから否定的な意味合いをもつ
あかごの手を捻ねる	非常に弱々しい赤ん坊の手は簡単にひねってしまえることのように、造作ない、簡単なことの例え
弱肉強食	弱い者が強い者の犠牲になるような、実力の違いが、そのまま結果に違いを生ずる闘争状態の世界。
同じ釜の飯を食う	ある程度の期間、他人同士が同じ家で起居を共にする、ないし、学校や職場・軍隊で生活を共にする。
百聞は一見に如かず	他人から何度聞いたところで、実際に自分の目で見る等体験して事実を知るという方法には及ばない。
怖いもの見たさ	怖いもの、恐ろしいものは、かえって好奇心がそそられ、興味本位で見たくなってしまうということ。
論より証拠	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。
安物買いの銭失い	安いものは品質が悪く、すぐに壊れて買い替える必要があるので、高いものを買うより損だということ。
一事（いちじ）が万事（ばんじ）	一つのことで全てが推測されるようす。普通は一つの悪い例を挙げて、そこから他の悪い様に敷衍する。
法の下に平等	(法律) 権利義務に関して法律上すべての人が平等に取り扱われなければならないという憲法上の原則。

TABLE 5.3: (continued)

Instruction	Sentence
疑心暗鬼を生ず	疑う心があると、何でもないことにまで恐ろしく感じられたり、疑いの気持ちを抱いたりするものである。
ボタンを掛け違える	手順を最初の方で間違えたために、当事者間での認識や考えに、その後ずっと続くようなずれが生まれる。
紆余曲折	道が曲がりくねって、真っ直ぐではないこと。物事の経緯やいきさつなどが、込み入った経過をたどること
七転び八起き	7 回転ぼうとも、それを超えて 8 回起き上がる気概で再起する、即ち、何度失敗しても立ち直るということ。
馬鹿の一つ覚え	覚えた事を、得意になって繰り返し何かにつけて言うこと。いつも同じことを言う人をあざけて言うことば。
切磋琢磨	学問などによって自分を磨いて、完成させること。また、同じ志を持つ人が互いに学問などを磨き合うこと。
意気投合	心が通じ合う事。意見が合うこと。考えなどがぴったりと一致して親しくなる事。「投合」は一致するの意味
大は小を兼ねる	大きい物であれば、小さい物の用途にも用いられる。余分に取っておけば、それに満たない物も補充できる。
意気消沈	元気をなくしてがっかりしている様子。意気込みが衰えて、しょげていること。「消沈」は「銷沈」とも書く。
自画自賛	自分で書いた絵に、自分で書いた賛（「讃」詩や文章）を添える事。自分で作った物や行為を自分で誉める事。
自暴自棄	破れかぶれで、自分を粗末に扱い、やけになること。「自暴」も「自棄」も「我が身を大事にしないこと」の意
棚から牡丹餅	（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。

Table 5.3: (continued)

Instruction	Sentence
一網打尽	（投網を一度投げてそこにいる魚をすべて取り尽すように）一度に関係者をことごとく捕らえ、罪に陥れること。
五里霧中	現在の状態が分からず、見通しや方針、手段の全く立たないことのたとえ。心が迷って、考えの定まらないこと。
油断大敵	失敗などの原因は、気のゆるみなど自らにあることが多いので、何事にも気を引き締めるべきであると言うこと。
試行錯誤	色々試みて、失敗を繰り返しながら目的に近づいていくというやり方。失敗を繰り返しながら解決法を探ること。
一騎当千	一人で千人ほどの敵に対し戦えるほど強いこと。力だけでなく、人並以上に優れた才能や経験の持ち主に対しても言う
鬼に金棒	強い者に更に強さが加わり、無敵となること。何も持たなくても強い鬼に、武器となる鉄棒を持たせるという意から。
呉越同舟	どんなに仲の悪い者同士であっても、共通の敵や災難にあっては、協力しこれを回避しようとするものであるということ
独活の大木	（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。
身から出た錆	自分の利益を増やすために、商品などの数量をごまかして伝えること。自分の年齢をごまかすことを指して使うことが多い
悠々自適	俗事に煩わされず、自分の思うままに静かに暮らすこと。（多くの場合、老年になって仕事から退いた人について言う。）
嵐の前の静けさ	大事や変事を前にして、奇妙に平穏であること、または、今は平穏であるが、それは大事や変事の前触れであるということ
地獄の沙汰も金次第	（「閻魔大王が下す地獄での判決も金次第では軽くもなる」ことから）世の中、金があれば何でも解決できるというたとえ。

TABLE 5.3: (continued)

Instruction	Sentence
かわいい子には旅をさせよ	厳しい経験を積むほど成長するため、かわいい子ほど敢えて辛い思いをさせよという意。昔の旅は辛いものだったことから。
図に乗る	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
千里の道も一歩から	千里の遠い所へ行くにも足元の第一歩から始まるの意味であって、大事を為すのにも小事を積み重ねることによって至るという譬え
情けは人の為ならず	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。
因果応報	良い行いをした人には良い報い、悪い行いをした人には悪い報いがある。つまり、やった行いに対しての報いが返ってくるという事
一蓮托生	（死後、浄土で同じ蓮華の上に生まれようという、日本の仏教上の思想から）物事の善悪や結果にとらわれず、行動を共にすること。
思い立ったが吉日	何かしようと決意したら、そう思った日を吉日としてすぐ取りかかるのが良いという意味。思い立つ日が吉日、思い立ったら吉日とも。
巨人の肩に立つ	先人の偉業にもとづいて仕事をすること。またそうすることにより、先人よりも能力が劣る人でも立派な業績をあげられるということ。
縁の下の力持ち	他人のために努力や苦労しても認知されない状況。転じて、人知れず陰で努力・苦労すること。またそのような人の例え。縁の下の舞。

TABLE 5.3: (continued)

Instruction	Sentence
犬も歩けば棒に当たる	犬がふらふら出歩くと、棒で殴られるような災難に遭ったりする。じっとしていれば良いのに、余計な行動を起こすべきでないとの戒め。
明鏡止水	(「明鏡」とは曇りのない澄んだ鏡、「止水」は静かに澄み、たまっている水）何の邪念も無く、静かに落ち着き澄み切っている心の状態。
ばつが悪い	その場・状況の文脈においてその特定の主体の行為・状況が不自然であるか恰好が悪いために、居づらい、気まずい、または場違いな様子。
只より高いものはない	一時的には、無料・無償であったり、非常に安価であったりするものは、後になって相応又はそれ以上の対価を支払うことになるものである
郷に入っては郷に従え	その土地（又は社会集団一般）に入ったら、自分の価値観と異なっていても、その土地（集団）の慣習や風俗にあった行動をとるべきである。
楽あれば苦あり	今は、安楽な思いをしていても、そのうち苦しいと思うときは来るものである、逆に、苦しいと思っているときもいつまでも続くものではない
嘘も方便	仏が衆生済度にあたっては、方便（手段）として嘘をつくこともある、ということから、大きな善行の前では、偽りも認められるということ。
虎穴に入らずんば虎子を得ず	(「虎が住んでいる穴に入らなければ虎の子を得ることは出来ない」ということから）時に危険な事柄をしなければ、成功することは出来ない。
壁に耳あり障子に目あり	隠し事をしようとしても、どこで誰が見たり聞いたりしているか分からないため、秘密・密談は漏れやすいものだから、注意しなさいという戒め。

Table 5.3: (continued)

Instruction	Sentence
起承転結	文章、特に4行から成る漢詩（近体詩）の絶句の構成方法。第1句が「起句」、第2句が「承句」、第3句が「転句」、第4句が「結句」である。
急がば回れ	危なくて短い道よりも安全で長い道を通ったほうが速く着くということから、物事は慌てずに着実に進めることが結果としてうまくいくということ。
捕らぬ狸の皮算用	（狸をまだ捕まえていないのに、その皮を売ったと考え、儲けの計算をすることから）手に入れていないものを当てにして、様々な計画を立てること。
諸行無常	仏教の基本的・哲学的な主張を表わす成句の一つで、「あらゆる物事（現象）は変化している。変化しない、固定的な物事は存在しない」という意味。
好きこそものの上手なれ	楽しんでやることによってうまくなるものであるということ、又は、あることに熟達するには、それを楽しめるようになることが肝要であるということ。
人を呪わば穴二つ	人を害すると、密かにやったつもりであっても、同じ仕打ちにあうことを覚悟すべきであるという事。転じて、安易に他人を害しようとすることを戒める
火に油を注ぐ	勢いが盛んなものに対し、さらに勢いを加えること。不本意なことについて用いられることが多い。「火に油」「火に油を加える」などとも表現される。
一富士二鷹三茄子	初夢で見ると縁起が良いといわれる三つのものを列挙した成句。一から順に、日本最高峰で霊峰の富士山、猛禽のタカ（鷹）、野菜のナス（茄子）を指す。
二兎追うものは一兎も得ず	うさぎを二兎同時に追いかけても、結局両方とも捕らえることはできない。二つのことを同時に成し遂げようとしても、結局どちらも失敗に終わるということ。

Table 5.3: (continued)

Instruction	Sentence
賽は投げられた	ユリウス･カエサルが、元老院体制に反旗を翻すべくルビコン川を渡る前に発した決断の台詞。転じて、もはや引き返せなくなる状態で、決断を促す際の台詞。
以心伝心	言葉を以ては伝えることのできない仏法上の真理を師から弟子に伝えること。言葉や文字を使わなくても、心と心で意思の疎通が出来る事。また、そう試みる事
雨降って地固まる	トラブルが発生したが、それが解決してしまうと、それが発生する前よりかえって良い状態になっていること、又は、往々にしてそういうものであるという達観。
河童の川流れ	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。
一期一会	人と人との出会いは、一生に一度のものであると心得、思い残すことの無いように接するべきであるとの教え。元は、茶道におけるもの。「一期」は一生の意味。
諸刃の剣	両刃（諸刃）の剣は、振り上げると自らも傷つける恐れがあることから、利益をもたらす可能性がある一方で、損失をもたらす危険性もはらんでいることのたとえ。
一心不乱	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。
臭い物に蓋をする	悪事や失敗、醜聞など、都合の悪いことが、他に漏れて世間に知られないように、根本的な解決をはかることなく、一時的にその場しのぎの方法で隠そうとすること。
明日は我が身	(他人事と思っていた事故や災難などが、明日には自分に降り掛かってくるかもしれないことから) 不幸な出来事が、いつ我が身にふりかかってくるかわからないこと。

TABLE 5.3: (continued)

Instruction	Sentence
弘法にも筆の誤り	書に優れている弘法大師であっても字を間違えることもあるということから、たとえその道の名人と呼ばれるような人間であっても、失敗をすることはあるという意味。
糠に釘	(糠に釘を打ち付けても手応えがないことから)手応えや効き目が全くないことの喩え。進んで、そのような手応えの無いものに働きかけることは無駄であることの戒め。
鬼の目にも涙	どんなに冷酷で無慈悲な性格の人間であっても、同情や憐れみを感じ涙を流すこともあるのだということ。強く恐ろしく見える鬼も泣くことがあるということから転じた。
破竹の勢い	勢いがあまりに激しく止めようのない状況、誰にも止められない快進撃を続けること。スポーツ選手の事績への言及や合戦、戦争での行軍を描写する際に用いることが多い。
石の上にも三年	(「石の上にじっと3年も座っていれば、石も暖まる」ということから)どんなに辛くても辛抱していれば、やがて、何らかの変化があって、好転の芽が出てくると言うこと。
遠くの親類より近くの他人	遠くに住んでいる親類縁者より、近くに住む隣人の方が、緊急を要するときには頼りになるものである、だから、日ごろ隣人とは良好な関係を築いておくべきであるということ
対岸の火事	(火事は大変な災いではあるが、川の向こう岸の火事はこちらへ燃え広がるおそれがないことから。)(他者にとっては大問題であっても)自分には関係なく、何の苦痛もないこと
井の中の蛙大海を知らず	「小さな井戸の中にいる蛙は、大きな海などの井戸の外にある世界のことを知らない」と言う意味から、自分の狭い知識にとらわれてしまい、物事の大局的な判断ができないこと。

Table 5.3: (continued)

Instruction	Sentence
良薬は口に苦し	（効き目のある薬が苦いように）いさめる言葉は、非難されているように聞こえ、素直に聞くことはできないものである。しかし、反省しその言に従うことが結局自分のためになる。
猫に小判	（猫に小判を与えても、その価値を知らない猫にとっては何の意味もないことから）どんな立派なものでも、価値がわからない者にとっては、何の値打ちもないものであるというたとえ。
阿鼻叫喚	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。
全ての道はローマに通ず	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。
唯我独尊	世界中において、人間のみが解脱することができるので尊いの意。釈迦が生まれたとき、一方の手は下（天下界）もう一方の手は上（天上界）を指し、7 歩歩いて辺りを顧みてから言ったとされる語。
塵も積もれば山となる	塵のように取るに足らない存在であっても、それが時間をかけて積もっていけば山のようになるように、些細な行動も、時間をかけて継続すると、やがて、思わぬ大きな結果につながるものであるということ。
船頭多くして船山に上る	指図する人が多過ぎるとかえって統率がとれず意に反した方向に物事が進んで行くことの意。「困難なことでも皆で力を合わせればできる」という解釈は誤り。なお、この場合の船頭は乗組員が複数いる場合の船長の意味。

TABLE 5.4: SUS questions in Experiment 2.

Question	Sentence
1	I think that I would like to use this system frequently
2	I found the system unnecessarily complex
3	I thought the system was easy to use
4	I think that I would need the support of a technical person to be able to use this system
5	I found the various functions in this system were well integrated
6	I thought there was too much inconsistency in this system
7	I would imagine that most people would learn to use this system very quickly
8	I found the system very cumbersome to use
9	I felt very confident using the system
10	I needed to learn a lot of things before I could get going with this system

Results of ROC curves

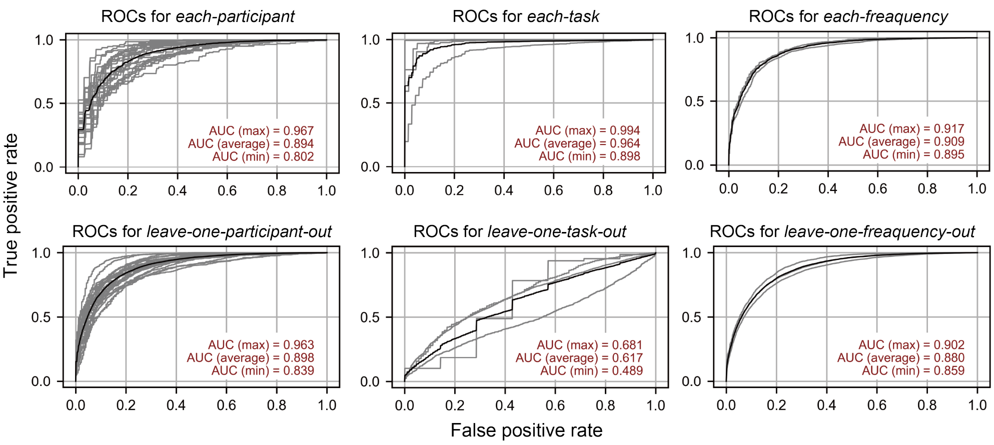

FIGURE 5.2: ROC curves for *each-XXX* and *leave-one-XXX-out*.

Other Author Publications

Journals

1. 礒本俊弥，山中祥太，志築文太郎．凝視後にジェスチャを行うという一連の操作を用いた意図しない操作に堅牢な視線に基づく操作手法．ヒューマンインタフェース学会論文誌 23 巻 1 号，pp. 5 − 18．ヒューマンインタフェース学会．2021 年発行．

2. 八箇恭平，礒本俊弥，志築文太郎，高橋伸．ターゲット内に両端が存在するスワイプジェスチャの設計と評価．コンピュータソフトウェア 37 巻 4 号 pp. 50 − 63．ソフトウェア科学会．2020 年発行．

Conference Papers

1. **Toshiya Isomoto**, Shota Yamanaka, Buntarou Shizuki. Reanalyzing Effective Eye-related Information for Developing User's Intent Detection Systems. In Proceedings of the 2023 ACM Symposium on Eye Tracking Research & Applications (ETRA 2023), Eyes4ICU, pp. 1-3, Tubingen, Germany. ACM, New York, NY, USA, May 2023..

2. **Toshiya Isomoto**, Shota Yamanaka, Buntarou Shizuki. Relationship between Dwell time and Model Human Processor for Dwell-based Image Selection. Proceedings of the ACM Symposium on Applied Perception 2021 (SAP 2021), Article 6, pp. 1-5. ACM, New York, NY, USA, virtual. Sep. 2021.

3. 礒本俊弥，山中祥太，志築文太郎．時間および範囲をもとに認識する凝視に基づく操作手法．第 28 回インタラクティブシステムとソフトウェアに関するワークショップ（WISS 2020）予稿集，Article 15，6 pages．ソフトウェア科学会，オンライン，2020 年 12 月．

4. **Toshiya Isomoto**, Shota Yamanaka, Buntarou Shizuki. Gaze-based Command Activation Technique Robust Against Unintentional Activation using Dwell-then-Gesture. Proceedings of Graphics Interface 2020 (GI 2020), Article 24, pp. 1-11, CHCCS/SCDHM, virtual, May 2020.

5. 八箇恭平，磯本俊弥，志築文太郎．シングルタッチジェスチャに対する片手操作手法の性能調査．第 24 回一般社団法人情報処理学会シンポジウム インタラクション 2020 予稿集，pp. 48-57，情報処理学会，東京，2020 年 3 月．

6. Kyohei Hakka, **Toshiya Isomoto**, Buntarou Shizuki, Shin Takahashi. Bounded Swipe: Swipe Gesture Inside a Target. Proceedings of the 30th Australian Conference on Computer-Human Interaction (OzCHI 2019), pp. 312-316, ACM, New York, NY, USA, Perth/Fremantle, Australia, December 2019.

7. 八箇恭平，磯本俊弥，志築文太郎．ターゲット内に両端が存在するスワイプジェスチャ．第 27 回インタラクティブシステムとソフトウェアに関するワークショップ（WISS 2019）予稿集，pp. 61-66，日本ソフトウェア科学会，長野，2019 年 9 月．

8. Toshiyuki Ando, **Toshiya Isomoto**, Buntarou Shizuki, and Shin Takahashi. Press & tilt: One-Handed Text Selection and Command Execution on Smartphone. Proceedings of the 30th Australian Conference on Computer-Human Interaction (OzCHI 2018), pp. 401-405, ACM, New York, NY, USA, Melbourne, Australia, December 2018.

9. **Toshiya Isomoto**, Toshiyuki Ando, Buntarou Shizuki, and Shin Takahashi. Dwell Time Reduction Technique using Fitts' Law for Gaze-Based Target Acquisition. Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA 2018), Article 26, pp. 1-7, ACM, New York, NY, USA, Warsaw, Poland, June 2018.

Demonstration and Poster Abstracts

1. **Toshiya Isomoto**, Shota Yamanaka, Buntarou Shizuki. Interaction Design of Dwell Selection Toward Gaze-based AR/VR Interaction. Proceedings of the 2022 ACM Symposium on Eye Tracking Research & Applications (ETRA 2022), Article No.: 39, 2 pages, June 2022.

2. **Toshiya Isomoto**, Shota Yamanaka, Buntarou Shizuki. Gaze-based Command Activation Technique using Two-level Stroke. Proceedings of the 2020 ACM CHI symposia on ASIAN CHI SYMPOSIUM: EMERGING HCI RESEARCH COLLECTION (Asian CHI Symposium 2020), 4 pages, ACM, New York, NY, USA, Hawaii, USA, April 2020.

3. Kyohei Hakka, **Toshiya Isomoto**, Buntarou Shizuki. One-Handed Interaction Technique for Single-Touch Gesture Input on Large Smartphones. Proceedings of the ACM Spatial User Interaction 2019 (SUI 2019), 2 pages, ACM, New York, NY, USA, New Orleans, USA, October 2019.

4. **Toshiya Isomoto**, Toshiyuki Ando, Buntarou Shizuki, and Shin Takahashi. Investigation of Midas-touches in Dwell Time Reduction Technique using Fitts' Law for Dwell-Based Target Acquisition. Proceedings of the 2019 ACM CHI symposia on ASIAN CHI SYMPOSIUM: EMERGING HCI RESEARCH COLLECTION (Asian CHI Symposium 2019), 8 pages, ACM, New York, NY, USA, Glasgow, UK, May 2019.

5. Toshiyuki Ando, **Toshiya Isomoto**, Buntarou Shizuki, and Shin Takahashi. One-handed Rapid Text Selection and Command Execution Method for Smartphones. Proceedings of Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA 2019), 6 pages, ACM, New York, NY, USA, Glasgow, UK, May 2019.

6. Ryosuke Takada, **Toshiya Isomoto**, Wataru Yamada, Hiroyuki Manabe, and Buntarou Shizuki. ExtensionClip: Touch Point Transfer Device Linking Both Sides of a Smartphone for Mobile VR Environments. Proceedings of Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18), ACM, New York, NY, USA. Montréal, Canada. April 2018.

7. **Toshiya Isomoto**, Akira Ishii, Shuta Nakamae, and Buntarou Shizuki. Target Selection Technique Using Space Below Cardboard VR Goggles. Proceedings of the 2018 ACM CHI symposia on ASIAN CHI SYMPOSIUM: EMERGING HCI RESEARCH COLLECTION (Asian CHI Symposium 2018), 8 pages, ACM, New York, NY, USA, Montréal, Canada, April 2018.

Misc.

1. 井口凌輔，横山海青，礒本俊弥，志築文太郎．WristRayFlick：手首からのレイによる VR 向けの片手かな文字入力手法．第 30 回インタラクティブシステムとソフトウェアに関するワークショップ（WISS2022）予稿集，2 pages，日本ソフトウェア科学会，宮城，2022 年 12 月．

2. 川口航平，礒本俊弥，志築文太郎，高橋伸．VR 向けの掌上における日本語フリック入力手法の提案．ヒューマンインタフェースシンポジウム 2019 論文集，pp. 676-682，ヒューマンインタフェース学会，京都，2019 年 9 月．

3. 礒本俊弥，山中祥太，志築文太郎．2 段階の視線移動を用いたコマンド実行手法．情報処理学会研究報告，Vol.2019-HCI-182 No.30，8 pages，情報処理学会，東京，2019 年 3 月．

4. 高田 崚介，礒本俊弥，山田 渉，真鍋 宏幸，志築文太郎．プロペラを用いた頭部装着型歩行牽引デバイス．第 23 回一般社団法人情報処理学会シンポジウム インタラクション 2018 予稿集，pp. 236-237，情報処理学会，東京，2019 年 3 月．

5. 礒本俊弥，安藤宗孝，志築文太郎．近接センサおよび照度センサを用いたスマートフォンベース HMD 向けの操作手法．第 26 回インタラクティブシステムとソフトウェアに関するワークショップ（WISS2018）予稿集，2 pages，日本ソフトウェア科学会，山梨，2018 年 9 月．

6. 安藤宗孝，礒本俊弥，志築文太郎，高橋伸．ソフトウェアキーボードのキーに基づく文字列操作手法．第 26 回インタラクティブシステムとソフトウェアに関するワークショップ（WISS2018），2 pages，日本ソフトウェア科学会，山梨，2018 年 9 月．

7. 安藤宗孝，礒本俊弥，志築文太郎，高橋伸．スマートフォンにおける傾きを利用した文字列操作手法．情報処理学会研究報告，Vol.2018-HCI-179，7 pages，情報処理学会，京都，2018 年 8 月

8. 礒本俊弥，安藤宗孝，志築文太郎，高橋伸．フィッツの法則に基づく視線を用いたターゲット選択システム．第 25 回インタラクティブシステムとソフトウェアに関するワークショップ（WISS2017），日本ソフトウェア科学会，山梨，2017 年 12 月．

9. 礒本俊弥，宮崎亮一．音声認識システムおよび顔の情報を用いた上肢不自由者のためのテキスト作成アプリケーションの開発．電子情報通信学会技術研究報告, 115(359), 金沢，pp. 59-62, 2015 年 12 月 11 日．

10. 礒本俊弥，宮崎亮一. 顔の方向情報を用いたテキスト編集補助アプリケーションの開発．第 21 回日本高専学会年会講演会講演論文集，徳山，P206，2015 年 8 月 29.