Description and Application of the Correlation between Gaze and Hand for the Different Hand Events Occurring During Interaction with Tablets.

Pierre Weill-Tessier

School of Computing and Communications

Lancaster University, UK

Thesis submitted for the degree of Doctor of Philosophy

January 2020

Abstract

People's activities naturally involve the coordination of gaze and hand. Research in Human-Computer Interaction (HCI) endeavours to enable users to exploit this multi-modality for enhanced interaction. With the abundance of touch screen devices, direct manipulation of an interface has become a dominating interaction technique. Although touch enabled devices are prolific in both public and private spaces, interactions with these devices do not fully utilise the benefits from the correlation between gaze and hand. Touch enabled devices do not employ the richness of the continuous manual activity above their display surface for interaction and a lot of information expressed by users through their hand movements is ignored.

This thesis aims at investigating the correlation between gaze and hand during natural interaction with touch enabled devices to address these issues. To do so, we set three objectives. Firstly, we seek to describe the correlation between gaze and hand in order to understand how they operate together: what is the spatial and temporal relationship between these modalities when users interact with touch enabled devices? Secondly, we want to know the role of some of the inherent factors brought by the interaction with touch enabled devices on the correlation between gaze and hand, because identifying what modulates the correlation is crucial to design more efficient applications: what are the impacts of the individual differences, the task characteristics and the features of the on-screen targets? Thirdly, as we want to see whether additional information related to the user can be extracted from the correlation between gaze and hand, we investigate the latter for the detection of users' cognitive state while they interact with touch enabled devices: can the correlation reveal the users' hesitation?

To meet the objectives, we devised two data collections for gaze and hand. In the first data collection, we cover the manual interaction on-screen. In the second data collection,

we focus instead on the manual interaction in-the-air. We dissect the correlation between gaze and hand using three common hand events users perform while interacting with touch enabled devices. These events comprise taps, stationary hand events and the motion between taps and stationary hand events. We use a tablet as a touch enabled device because of its medium size and the ease to integrate both eye and hand tracking sensors. We study the correlation between gaze and hand for tap events by collecting gaze estimation data and taps on tablet in the context of Internet related tasks, representative of typical activities executed using tablets. The correlation is described in the spatial and temporal dimensions. Individual differences and effects of the task nature and target type are also investigated.

To study the correlation between gaze and hand when the hand is in a stationary situation, we conducted a data collection in the context of a Memory Game, chosen to generate enough cognitive load during playing while requiring the hand to leave the tablet's surface. We introduce and evaluate three detection algorithms, inspired by eye tracking, based on the analogy between gaze and hand patterns. Afterwards, spatial comparisons between gaze and hands are analysed to describe the correlation. We study the effects on the task difficulty and how the hesitation of the participants influences the correlation. Since there is no certain way of knowing when a participant hesitates, we approximate the hesitation with the failure of matching a pair of already seen tiles. We study the correlation between gaze and hand during hand motion between taps and stationary hand events from the same data collection context than the case mentioned above. We first align gaze and hand data in time and report the correlation coefficients in both X and Y axis. After considering the general case, we examine the impact of the different factors implicated in the context: participants, task difficulty, duration and type of the hand motion.

Our results show that the correlation between gaze and hand, throughout the interaction, is stronger in the horizontal dimension of the tablet rather than in its vertical dimension, and that it varies widely across users, especially spatially. We also confirm the eyes lead the hand for target acquisition. Moreover, we find out that the correlation between gaze and hand when the hand is in the air above the tablet's surface depends on where the users look at on the tablet. As well, we show that the correlation during eye and hand during stationary hand events can indicate the users' indecision, and that while the hand is moving, the correlation depends on different factors, such as the degree of difficulty of the task performed on the tablet and the nature of the event before/after the motion.

Declaration

This thesis is a presentation of my original research work. No part of this thesis has been submitted for another degree or qualification. The work was done under the guidance of Professor Hans Gellersen at Lancaster University.

Author: **Pierre Weill-Tessier**

Acknowledgements

For the patience and the guidance I have received during my PhD, I first would like to thank my supervisor, Professor Hans Gellersen.
For their valuable feedback and interest in my work, I would like to express my gratitude to Doctor Parisa Eslambolchilar and Doctor Keith Cheverst, as well as Doctor Abe Karnik and Doctor Matthew Broadbent.
Thanks also to all the persons that I have worked with, or simply met, at the School of Computing and Communications (especially fellow students and staff of the EIS/Interaction Lab Group) for their advices and their friendship.

I am also grateful for the support given by my friends from the Confucius Institute, and those from outside Lancaster: in France, in York, in Loughborough, in Pinneberg, in Melbourne, and in China.

Finally, I would like to express my sincere gratitude to my family who helped me so much and who always placed confidence in me...

Publications

This work has been published in peer-reviewed publications at conferences. Below are the references of these publications.

- Pierre Weill-Tessier, Jayson Turner, and Hans Gellersen. How do you look at what you touch? In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications - ETRA '16*, ETRA '16, pages 329–330, New York, New York, USA, 2016. ACM Press
  (in Chapter 4)

- Pierre Weill-Tessier and Hans Gellersen. Touch input and gaze correlation on tablets. In Ireneusz Czarnowski, Robert J. Howlett, and Lakhmi C. Jain, editors, *Intelligent Decision Technologies 2017: Proceedings of the 9th KES International Conference on Intelligent Decision Technologies (KES-IDT 2017) – Part II*, pages 287–296. Springer International Publishing, Cham, 2018
  (in Chapter 4)

- Pierre Weill-Tessier and Hans Gellersen. Correlation between gaze and hovers during decision-making interaction. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications - ETRA '18*, ETRA '18, pages 1–5, New York, New York, USA, 2018. ACM Press
  (in Chapter 5 - Sections 5.4 and 5.5)

Contents

Abstract i

Declaration iii

Acknowledgements iv

Publications v

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Methodology . . . 4
- 1.3 Contributions . . . 7
- 1.4 Thesis Structure . . . 8

2 Background 10
- 2.1 Tracking Gaze, Tracking Hands . . . 10
  - 2.1.1 Overview of the Eye Tracking . . . 11
    - 2.1.1.1 Description of the Eye and Visual System . . . 11
    - 2.1.1.2 Vision Characteristics . . . 12
    - 2.1.1.3 Tracking the Eyes . . . 13
  - 2.1.2 Overview of Hand Tracking . . . 15
    - 2.1.2.1 Hand Biomechanism Involved in Aiming and Reaching . . . 15
    - 2.1.2.2 Hand Tremor . . . 16
    - 2.1.2.3 Tracking the hand . . . 16
- 2.2 Correlation between Gaze and Hand . . . 19
  - 2.2.1 Gaze and Hand Coordination in Aiming or Reaching . . . 20
  - 2.2.2 Mouse and Gaze . . . 22
  - 2.2.3 Correlation between Gaze and Hand Movements . . . 23
- 2.3 Study Contexts . . . 24
  - 2.3.1 Internet Based Tasks in Research . . . 24
  - 2.3.2 Decision Making Study Activities . . . 27
- 2.4 Human-Computer Interaction Applications . . . 28
  - 2.4.1 Applications from Gaze and Hand Modalities . . . 29
    - 2.4.1.1 Using Gaze and Hand Modalities Independently . . . 29
    - 2.4.1.2 Using the Correlation between Gaze and Hand Modalities 30
  - 2.4.2 Applications from Manual Input above the Interactive Surface . . . 31
  - 2.4.3 Hesitation Detection . . . 33
  - 2.4.4 Intelligent and Adaptive User Interfaces . . . 34

3 Study Setup 37
- 3.1 Eye Tracker . . . 37
  - 3.1.1 Eye Tracker Installation . . . 37
  - 3.1.2 Eye Tracker Calibration . . . 39

3.1.3 Eye Tracker Data Interpretation . . . 41
3.1.4 Eye Tracker Limitations . . . 42
3.2 On-Screen Hand Gesture Detection . . . 43
3.2.1 Raw Manual Input Data . . . 43
3.2.2 Third Party Gesture Detection . . . 46

4 Correlation between Gaze and Tap 49
4.1 Introduction . . . 49
4.2 Pilot study . . . 50
4.2.1 Introduction . . . 50
4.2.2 Method and Apparatus . . . 50
4.2.3 Results . . . 51
4.2.3.1 Completion Time . . . 51
4.2.3.2 Distance Error . . . 52
4.2.3.3 Failed Tap Attempts . . . 53
4.2.4 Conclusion . . . 54
4.3 Study Design . . . 55
4.3.1 Data Collection Context . . . 55
4.3.1.1 Search Task . . . 55
4.3.1.2 Shopping Task . . . 56
4.3.1.3 Game Task . . . 56
4.3.2 Study Protocol . . . 57
4.3.3 Study Architecture and Data Collection . . . 58
4.3.4 Data Collection Content Overview . . . 62
4.4 Fixations around Tap Moment . . . 63
4.4.1 Spatial Distribution of All Fixations at Tap . . . 63
4.4.2 Number of Fixation Before and After Tap . . . 64
4.4.3 Relationship between Spatial and Temporal Distribution of the Fixations around Tap . . . 66
4.4.4 Impact of Individuals, Tasks and Target Types . . . 67
4.4.4.1 Across Participants . . . 67
4.4.4.2 Across Tasks . . . 68
4.4.4.3 Across Target Types . . . 68
4.5 A Specific Fixation: $F_{Closest}$ . . . 70
4.5.1 $F_{Closest}$ General Characteristics . . . 70
4.5.2 Spatial Distribution of $F_{Closest}$ Relative to Taps . . . 71
4.5.3 Relationship between $F_{Closest}$ and Tap . . . 74
4.6 Typing . . . 75
4.7 Discussion . . . 81
4.8 Conclusion . . . 82

5 Correlation between Gaze and Stationary Hand Event 84
5.1 Introduction . . . 84
5.2 System Design . . . 85
5.2.1 Content . . . 85
5.2.2 Context . . . 86
5.2.3 Apparatus . . . 86
5.2.4 Protocol . . . 87
5.2.5 Participants . . . 88
5.2.6 Eye Movements Classification . . . 89
5.2.7 Hand Events Classification . . . 89
5.3 Stationary Hand Events Detection . . . 90

5.3.1 Data Preparation . . . 90
5.3.1.1 Ground Truth . . . 90
5.3.1.2 Hovers and Dwell Definitions . . . 92
5.3.2 Algorithms Presentation . . . 92
5.3.3 Algorithms Performance . . . 94
5.3.3.1 Dwells and Hovers . . . 95
5.3.3.2 Dwells Only . . . 97
5.3.3.3 Hovers Only . . . 101
5.3.3.4 Test . . . 102
5.4 Relationship between Gaze and Stationary Hand Events . . . 102
5.5 Indecision and Gaze/Stationary Hand Events Relationship . . . 106
5.6 Discussion . . . 108
5.7 Conclusion . . . 110

6 Correlation between Gaze and Hand in Motion 112
6.1 Introduction . . . 112
6.2 Data Preparation . . . 113
6.2.1 Context . . . 113
6.2.2 Method . . . 113
6.3 Results . . . 114
6.3.1 Overall Correlation . . . 115
6.3.2 Correlation per Participant . . . 116
6.3.3 Correlation per Game Level . . . 118
6.3.4 Impact of Time on the Correlation . . . 119
6.3.5 Impact of the Motion Type . . . 120
6.3.6 Spatial Difference between Gaze and Hand . . . 122
6.4 Discussion . . . 123
6.5 Conclusion . . . 125

7 Discussion 127
7.1 Limitations . . . 128
7.1.1 Apparatus . . . 128
7.1.2 Tasks . . . 128
7.1.3 Data . . . 130
7.2 Further Research . . . 130
7.2.1 Human Factors and Sensors . . . 130
7.2.2 Contexts . . . 133
7.2.3 Replication . . . 134
7.2.4 Applications . . . 134

8 Conclusion 136

Bibliography 139

Appendix 163
A Gaze and Tap Correlation Study Material . . . 163
B Gaze and Hovers Correlation Memory Game Details . . . 168
C Gaze and Hovers Correlation Other Results . . . 171
D Gaze and Hand Motion Correlation Other Results . . . 182

List of Figures

1.1 Breakdown of the typical hand events observed during tablet interaction. . 5
1.2 Diagram of the research questions' coverage over the different hand events studied in the thesis. 7

2.1 Structure of the human eye. 11

3.1 Usual remote eye tracker configuration. 38
3.2 Tobii X2-60 configuration tool. 39
3.3 Tobii EyeX configuration via API. 40
3.4 Eye tracker precision and accuracy. 41
3.5 Code snippet showing how to retrieve and register the digitizer with the Raw Input API. 44
3.6 Code snippet showing how to read the HID reports data of the digitizer with the Raw Input API. 45
3.7 Code snippet showing how the raw input is processed by Sparsh UI. . . . 46
3.8 Code snippet showing what selected gestures are interpreted by Sparsh UI. 47
3.9 Code snippet showing the writing of on-screen gestures log files with Sparsh UI. 48

4.1 Pseudo-random sequence order of the pilot study targets. 51
4.2 Completion time (per target, per condition). 52
4.3 Distance error (per target, per condition). 53
4.4 Mean failed tap attempts (per target, per condition). 54
4.5 Search task. 56
4.6 Shopping task. 57
4.7 Game task. 58
4.8 Eye tracker and stand configuration. 59
4.9 Browser's navigation bar (truncated). 60
4.10 System architecture. 61
4.11 Spatial distribution of the fixations relative to the tap position (2 seconds around the tap). 63
4.12 Spatial distribution of the fixations relative to the tap position (2 seconds around the tap). 64
4.13 Distance percentages of the fixations relative to the tap position at different time windows relative to the tap moment (50 ms wide). 65
4.14 Number of fixations strictly before the tap moment. 65
4.15 Number of fixations at and after the tap moment. 66
4.16 Fixation start moment vs. fixation distance (relative to tap moment/position). 67
4.17 Fixation start moment vs. fixation distance (relative to tap moment/position, per participant). 69
4.18 Fixation start moment vs. fixation distance (relative to tap moment/position, per task). 70

4.19 Fixation start moment vs. fixation distance (relative to tap moment/position, per target type). 70
4.20 $F_{Closest}$'s start moment histogram and quartiles. 71
4.21 $F_{Closest}$'s distance histogram and quartiles. 72
4.22 $F_{Closest}$'s mean and median positions around tap position. 72
4.23 $F_{Closest}$'s mean position around tap position (per participant). 73
4.24 $F_{Closest}$'s median position around tap position (per participant). 73
4.25 $F_{Closest}$'s mean and median positions around tap position (per task). . . . 74
4.26 $F_{Closest}$'s mean and median positions around tap position (per target type). 74
4.27 Fixations (blue) and taps (black) during a typing sequence. 75
4.28 $F_{Closest}$ associated with taps (red) and taps (black) during a typing sequence. 76
4.29 $F_{Closest}$ associated with taps (red) and taps (black) during a typing sequence (poor alignment). 77
4.30 Average $F_{Closest}$ start moment before keyboard tap per typing skill level. . 78
4.31 Average distance between $F_{Closest}$ and keyboard tap's locations per typing skill level. 79
4.32 Average horizontal distance between $F_{Closest}$ and keyboard tap's locations per typing skill level. 80
4.33 Average vertical distance between $F_{Closest}$ and keyboard tap's locations per typing skill level. 80

5.1 Data visualisation for the feasibility study for one participant. 87
5.2 System used during the data collection of stationary hand events. 88
5.3 Screenshots of different hand moving trend sequences during interaction. . 91
5.4 Projection (P) of the hand on the tablet from the user's eye perspective. . 92
5.5 IDT Precision-Recall space for the IDT algorithm (grouping by Tt). . . . 96
5.6 Precision-Recall space for the IDT algorithm (grouping by St). 96
5.7 F1 score for the different combinations of thresholds of the IDT algorithm. 97
5.8 Precision-Recall space for the IDTE algorithm (grouping by Tt, dwells only). 98
5.9 Precision-Recall space for the IDTE algorithm (grouping by St, dwells only). 99
5.10 F1 score for the different combinations of thresholds of the IDTE algorithm (dwells only). 99
5.11 Precision-Recall space for the IVT algorithm (grouping by Tt, hovers only).100
5.12 Precision-Recall space for the IVT algorithm (grouping by St, hovers only).100
5.13 F1 score for the different combinations of thresholds of the IVT algorithm (hovers only). 101
5.14 Relative median position between gaze and stationary hand events per tile. 103
5.15 Relative median position between gaze and stationary hand events per tile for *left-handed* participants. 104
5.16 Relative median position between gaze and stationary hand events per tile for *right-handed* participants. 105
5.17 Relative median position between gaze and hover per tile. 106
5.18 Median distance between gaze and stationary hand event positions. 107
5.19 Median distance between gaze and stationary hand event positions. 108

6.1 2D distribution of the gaze and hand data. 115
6.2 Participant's Spearman correlation coefficients boxplots per axis. 116
6.3 Distribution of the Spearman correlation coefficients over participants for X and Y axis (all levels). 117
6.4 Spearman's correlation coefficients relationship over the participants for X and Y axis (all levels). 117

6.5 Spearman's correlation coefficients relationship over the game levels for X and Y axis. . . . 118
6.6 Hand motion duration distribution. . . . 119
6.7 Spearman's correlation coefficients relationship over the motion duration ranges for X and Y axis. . . . 119
6.8 Hand motion type distribution. . . . 121
6.9 Spearman's correlation coefficients relationship over the motion types for X and Y axis. . . . 122
6.10 Mean difference between gaze and hand during hand motion per tile (based on gaze location). . . . 123

7.1 Example of the hand leaving the interaction space as a potential expression of frustration, hesitation or reflection. . . . 130

A.1 Questionnaire submitted at the end of the study. . . . 165
A.2 Snippet of the JavaScript code injection on the webpages on the emulated browser. . . . 166
A.3 Flyer of the data collection participation (related to Chapter 4). . . . 167
B.1 Schematic disposition of the apparatus elements. (sideways view) . . . 168
B.2 Schematic disposition of the apparatus elements. (top view, alignment is suggested by the dotted line) . . . 168
B.3 Demonstration Version of the Memory Game. . . . 169
B.4 Level 1 of the Memory Game. . . . 169
B.5 Level 2 of the Memory Game. . . . 169
B.6 Level 3 of the Memory Game. . . . 170
B.7 Flyer of the data collection participation (related to Chapters 5 and 6). . . . 170
C.1 Precision-Recall space for the IDTE algorithm (grouping by Tt). . . . 171
C.2 Precision-Recall space for the IDTE algorithm (grouping by St). . . . 171
C.3 Precision-Recall space for the IVT algorithm (grouping by Tt). . . . 172
C.4 Precision-Recall space for the IVT algorithm (grouping by St). . . . 172
C.5 F1 score for the different combinations of thresholds of the IDTE algorithm. 173
C.6 F1 score for the different combinations of thresholds of the IVT algorithm. 173
C.7 IDT Precision-Recall space for the IDT algorithm (grouping by Tt, dwells only). . . . 174
C.8 Precision-Recall space for the IDT algorithm (grouping by St, dwells only). 174
C.9 Precision-Recall space for the IVT algorithm (grouping by Tt, dwells only). 175
C.10 Precision-Recall space for the IVT algorithm (grouping by St, dwells only). 175
C.11 F1 score for the different combinations of thresholds of the IDT algorithm (dwells only). . . . 176
C.12 F1 score for the different combinations of thresholds of the IVT algorithm (dwells only). . . . 176
C.13 IDT Precision-Recall space for the IDT algorithm (grouping by Tt, hovers only). . . . 177
C.14 Precision-Recall space for the IDT algorithm (grouping by St, hovers only). 177
C.15 Precision-Recall space for the IDTE algorithm (grouping by Tt, hovers only). . . . 178
C.16 Precision-Recall space for the IDTE algorithm (grouping by St, hovers only). 178
C.17 F1 score for the different combinations of thresholds of the IDT algorithm (hovers only). . . . 179
C.18 F1 score for the different combinations of thresholds of the IDTE algorithm (hovers only). . . . 180
C.19 Median Distance between gaze and stationary hand event (per participant). 181

List of Tables

4.1 Mean completion time (per participant, per condition). 52
4.2 Mean distance error (per participant, per condition). 53
4.3 Mean failed tap attempts (per participant, per condition). 54
4.4 Fixation start moment vs. fixation distance (minima, per participant). . . 68
4.5 Fixation start moment vs. fixation distance (minima, per task/target type) 69
4.6 Pearson correlation between the $F_{Closest}$/tap distance and the target position/size. 75

5.1 Stationary hand events (hovers and dwells) number and round duration percentage for the validation subset. 89
5.2 Best F1 scores per algorithms (dwell and hover). 97
5.3 Best F1 scores per algorithms (dwell only). 98
5.4 Best F1 scores per algorithms (hover only). 101
5.5 Testing set results. 102
5.6 Testing set results for best F1 values. 102
5.7 Classification of the hovers. 103
5.8 Median distance between gaze and stationary hand event positions (per game level). 107

6.1 Pair-wise Z-score for the Spearman correlation coefficients between gaze and hand comparison per game level. 118
6.2 Pair-wise Z-scores for the Spearman correlation coefficients between gaze and hand comparison (X axis) per duration range. 120
6.3 Pair-wise Z-scores for the Spearman correlation coefficients between gaze and hand comparison (Y axis) per duration range. 120

A.1 Questions of the Search Task. 163
A.2 Suggestion of Mock-up Personal Data for the Shopping Task. 163
A.3 Source and Target Articles (2 rounds) for the Game Task. 164
D.1 Pair-wise Z-scores for the Spearman correlation coefficients between gaze and hand comparison (X axis) per hand motion type. 182
D.2 Pair-wise Z-scores for the Spearman correlation coefficients between gaze and hand comparison (Y axis) per hand motion type. 183

1

Introduction

1.1 Motivation

Until recently, multimodal interaction with computer devices was secluded to the environment of the research labs. The miniaturisation and the reliability of body sensors and human activity recognition devices permitted a deployment to public reach. Therefore, multimodal interaction constitutes a trendy topic in the field of Human-Computer Interaction. Among the different possible body parts tracked by systems, gaze and hand hold a special place since they are, as mentioned in studies from psychology and neuroscience, considerably involved in human ordinary activities and work in a complementary fashion [121, 156, 169, 199].

In activities related to computer interaction, researchers demonstrated the role gaze and hand can play together in the usability of the computing devices. Early Human-Computer Interaction semi-theoretical work presented the concepts and expected promises of eye tracking interaction [103, 104, 123], later put to the test in more recent works for practical applications with combined gaze input and manual input [158, 159, 160, 161, 162, 178, 179, 180, 190, 191, 192, 222, 223]. Initially, the correlation between gaze and hand while interacting with computing devices employed the mouse, indirect manual input, to replace the hand, for example to predict the user's click. However, touch enabled devices, such as tablets, kiosks and mobiles, offer a better representation of the actual correlation between gaze and hand since the hand is directly involved into the interaction process. Nevertheless, research work dealing with hand and gaze interaction on touch en-

abled devices discard the description and exploitation of the natural correlation between gaze and hand. Instead, they introduce *new* interaction techniques in which gaze and hand collaborate in separate spaces. The correlation between gaze and hand on touch enabled devices is therefore non existent in literature, despite the potential improvement it could bring to the interaction with those devices. Indeed, we encompass interaction can benefit from techniques based on the correlation between gaze and hand on touch enabled devices for two reasons: (1) because research in Human-Computer Interaction already demonstrates concrete example of how the correlation between gaze and hand (via the mouse) provide applications with desktop computers, and (2) because this work relies on a natural and unconscious behaviour of the users who, shall applications be in place in the future, will not necessitate any learning nor constrain to experience these applications.

Hand and eye coordination when interacting with touch-enabled devices cognitively joins or differs from the coordination implicated in tools handling on several points. According to Vaesen [195], humans' handedness brought increased dexterity. Some tools may be designed for right-handed usage, as more than 85 % of humans are right-handed, and therefore penalise left-handed people. In computer direct interaction such as the one found with touch-enabled devices, this problem is partly suppressed (the user is free to interact with her dominant hand, but the application layout may be designed for right-handed people): no cognitive effort is required in this manual interaction to adapt the dexterity to the handedness. More importantly, direct interaction bypass the '*function representation*" Vaesen assimilates with the human representation of tools: when interacting with touch-screen, the hand is not used as an extension of the body - it **is** the body. However, the "*casual reasoning*", "*social learning*" and "*cumulative culture*" Vaesen exposes in his paper can be met with touch-enabled devices. Although tapping on touch-enabled surfaces is considered as natural and intuitive [20], the gestures that allow a complete interaction (such as zooming) are not, and may require (easy and fast) learning from observation.

In this thesis, we aim at exploring the correlation between hand and gaze on touch enabled devices, for the different types of hand events occurring during the interaction with a tablet, in *natural* activities. In more details, the research work presented by this thesis follows the following objectives:

1. We first want to give a plain description of this correlation: how does the hand and the eye behave with one another in the temporal and in the spatial dimensions?

That objective's outcome adds up to the current understanding of how the human central nervous control system manages the hand/eye correlation in general - since the direct manual input, found in touch enabled devices usability, presents a similarity with psychological studies already analysing how gaze and hand behave in target selection.

2. The general view of the correlation between gaze and hand we tackle at first logically leads to the questioning of how some basic factors found in the natural interaction with touch devices impact the correlation : how, if at all, individual differences, the nature/complexity of the tasks, or the targets impact the correlation between gaze and hand? We limit the scope of the factors to those three for their implicit occurrence in the interaction and the ease for measurement. Other possible factors, such as the state of arousal, despite being implicit would be measured by other means that we did not include in our study (for example galvanic skin response).

3. In the field of Human-Computer Interaction, understanding how gaze and hand work together serves as the foundation for building up applications. For the direct manual input modality our thesis relates to, we want to make use of the correlation between gaze and hand in order to provide the touch enabled devices a way to assess the user's cognitive state: does the correlation between gaze and hand show any characteristics that indicate the user's hesitation/indecisiveness during decision making activities on a touch enabled device? We decide to direct our research towards this type of application to contribute to the *human centred* Human-Computer Interaction (Intelligent Human-Computer Interaction, sometimes referred as HCI^2 [152]). Intelligent Human-Computer Interaction promises enhancements in, for instance, collecting feedback on the interaction with applications to improve their design, monitoring users' activity to assess their medical condition, or personalising evolving interfaces.

Natural interaction ought to be maintained to address all to points mentioned above. The rationale behind this requirement essentially emerges from the very meaning of studying the correlation between gaze and hand during touch enabled devices interaction: we focus on an innate human behaviour.

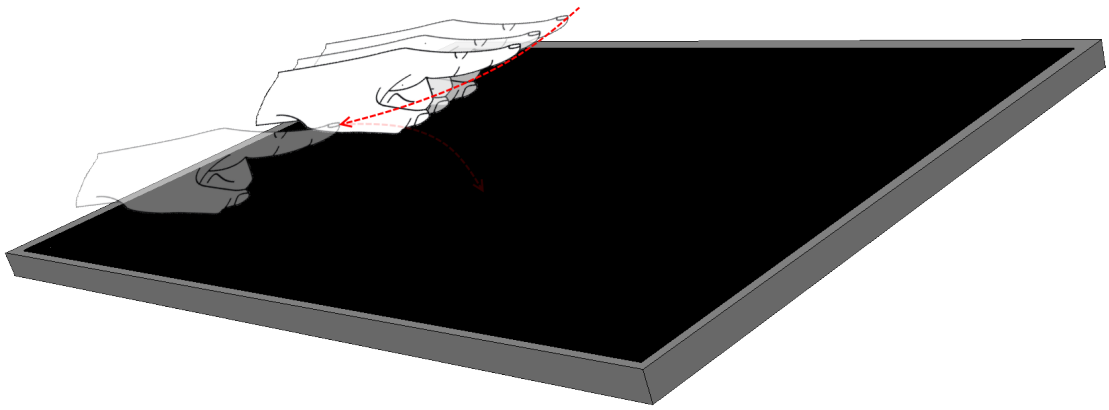
Overall, we considered these objectives for our willingness to provide a baseline understanding of the human behaviour while using touch-enabled devices, such as public kiosks on which personalisation of the device is limited by short interaction and perma-

nent changes of users. The research questions tackled in the this thesis are therefore related to answer this general question: how the correlation between gaze and hand can be exploited by adaptive interfaces to provide intelligent output towards the users' behaviours. We limit our exploration of the answer with the factors mentioned in Objective (2) and with one type of user's behaviour: hesitation, in Objective (3). Nevertheless, we will address further exploration of the answer in the general discussion (Chapter 7) of the thesis.

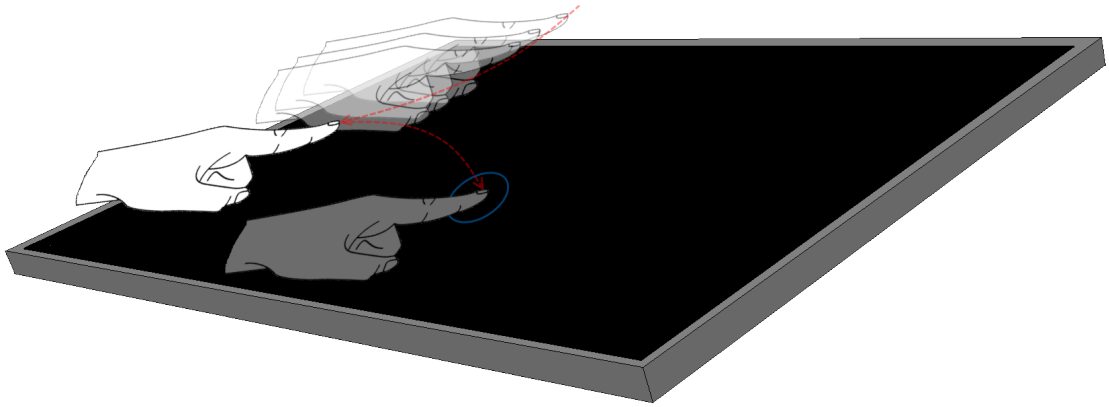
1.2 Methodology

Among the possible touch enabled devices to study the correlation between gaze and hand on (mobile phones, tablets, tabletop large display...), we selected a tablet. The reason behind this selection is the average size of such device, the possibility to simulate a device used either at home or in public spaces (kiosks) and the ease for implementing a study context.

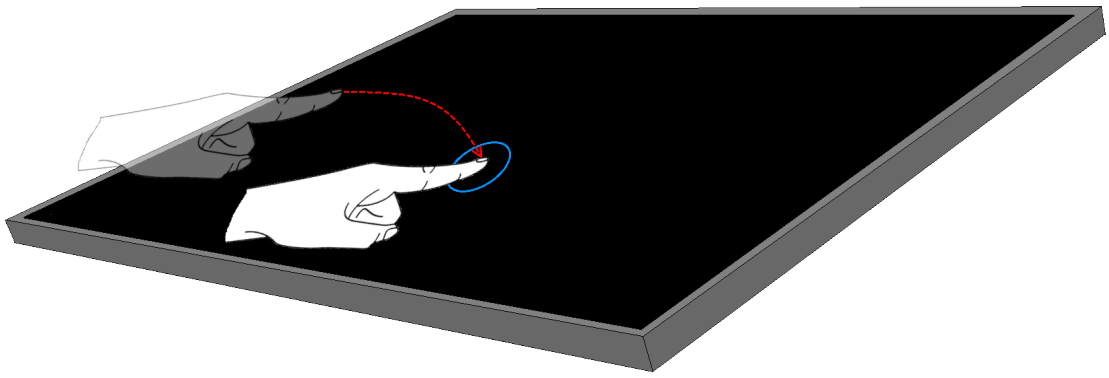
Earlier work on hand and gaze correlation in computer interaction used the mouse, since it was the common and standard manual indirect input device in place with desktop computers. The mouse is not permanently active: Chen et.al, in [33] considered the different *activities* of the mouse in web browsing, and identified the "nowhere" area where the mouse is left inactive. The information conveyed from the mouse in the "left inactive" state is rather limited: it indicates the users are *not* engaged in a clicking action, but shows nothing on how they are actually *still* involved into the interaction with the desktop device. Similarly, when interacting with a touch enabled device that detects tapping as the sole manual input, an interruption in the interaction is perceived any time the hand leaves the device's surface. But the hand actually does not always become inactive during this interruption (of tapping), contrary to the counterpart idleness of the mouse when the user is not engaged in a clicking process. The hand in this situation is evolving in the volume above the surface, moving or marking "pauses", which can inform about the user's cognitive state in a much richer way than the idle mouse [3]. That is why, in this thesis, we choose to analyse the correlation between gaze and hand according to the different hand events that commonly occurs during interaction with tablets: taps, stationary hand events and hand motion in between the two first events.



(a) The hand is moving.



(b) The hand is in stationary position (hover or dwell).



(c) The hand is performing a tap.

Figure 1.1: Breakdown of the typical hand events observed during tablet interaction.

A typical breakdown of these continuous hand events is given in Figure 1.1. Often while preparing a tap, the hand is moving (Figure 1.1a) until it stops for a short period of time (Figure 1.1b). Then, the hand either performs a tap (Figure 1.1c) or moves again to another destination.

Work related to the tapping part of the hand events, illustrated by Figure 1.1c, is detailed in Chapter 4. This hand event bears the closest resemblance with the clicking action studied in other research activities related to the desktop computer mouse/gaze relationship. The support for this work is a data collection on taps while performing Internet related tasks on a tablet. The tasks are designed to echo the activities users commonly performed on a tablet, and to ensure as much as possible a natural interaction while generating enough taps. The apparatus chosen for the data collection has been decided after validation with a pilot study. We conduct an exploratory approach in the study of the correlation between gaze and hand for taps.

The next hand event type we focus on is the *stationary hand event*, illustrated by Figure 1.1b. Stationary hand events are categorised in two natures: hovers (when the hand is strictly above the surface) or dwells (when the hand is left outside the interaction volume). To detect them, we develop and evaluate several algorithms based on eye tracking techniques. The reason behind this choice is the observation of an analogy between the stationary hand events and the eye fixations. The data is collected from another context than the previous work: we designed our own stand to combine a tablet, an eye tracker and a hand tracker, and participants played a Memory Game, so we could stimulate their cognitive activity and generate stationary hand events, while offering an experience to the participants that can still be assimilated with a natural activity they would perform on a tablet. The correlation between gaze and stationary hand events is explored in Chapter 5, which includes the detection algorithms' presentation in Section 5.3. In this chapter, we also study the differences in the correlation between gaze and hand induced by the decisiveness of the participants. The decisiveness in our study is approximated by the success of the game's tile pair matching.

The last hand event we analyse is the connection with the first two events mentioned above: when the hand moves between taps, hovers and dwells. Chapter 6 explains how the correlation between gaze and hand is computed. The context for the data collection is the same as in Chapter 5. We evaluate how strong the correlation is between gaze and hand during hand motion, and we show how the correlation was impacted by application factors (such as the degree of difficulty, the interaction area of the surface) and human

factors (individual differences, duration of the motion type of the motion).

Notwithstanding the three chapters mentioned in this section are all dealing with the correlation between gaze and hand, the points of view to tackle the analysis are not the same. In Chapter 4, the study unit is the tap and we analysed how gaze behaves in relation to it, whereas in Chapter 5 the study unit is the area of gaze and we analysed how the hand behaves respecting it. For Chapter 6, the point of view is different again since we analyse gaze and manual movements altogether: no specific variable serves as a reference to study the other.

We address the research questions according to the hand event as summarised by a summary diagram (Figure 1.2):

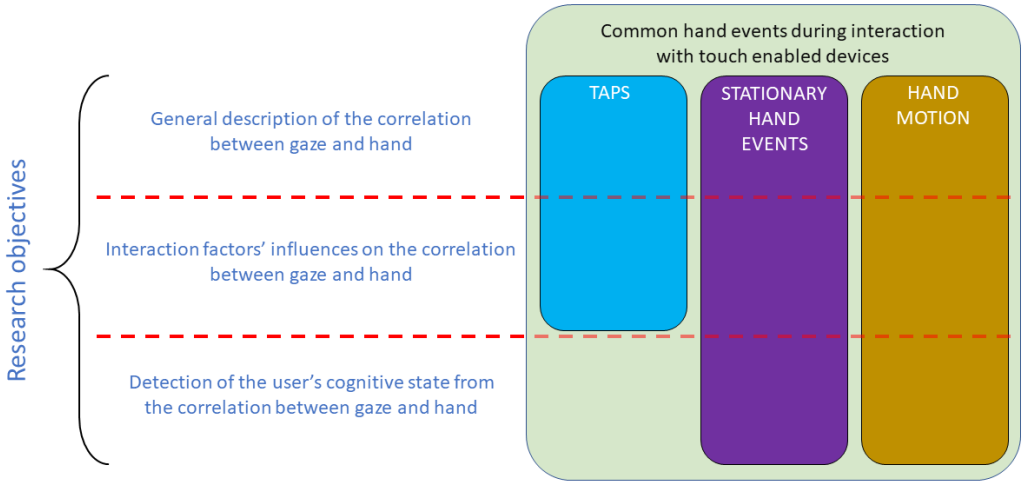


Figure 1.2: Diagram of the research questions' coverage over the different hand events studied in the thesis.

1.3 Contributions

This thesis brings the following contributions to the field of Human-Computer Interaction:

- The description of the correlation between hand and gaze during interaction with touch enabled devices, sectioned by the different basic hand events we notified from observation.

- The proposal of two study contexts for a gaze and hand data collection: one context contains Internet based activities (data collected: gaze/tap) and the other context contains a cognitive load activity (a Memory Game, data collected: gaze/hand movement).
- The proposal and the evaluation of three algorithms to extract stationary hand events from hand motion data, based on an analogy between gaze and hand movement patterns.
- Elements to implement an example of Intelligent Human-Computer Interaction, by retrieving hesitation based on the correlation between hand and gaze on touch enabled devices, and by evaluating a task's difficulty from the changes in the correlation between gaze and hand when the hand is moving between taps and stationary hand events.

1.4 Thesis Structure

The thesis is articulated around the following chapters:

- **Chapter 2 - Background**: references the milestones and techniques dealing with eye tracking and hand tracking, and presents key understandings of the hand and gaze coordination in aiming or reaching, between the mouse and gaze, and during hand motion, because these actions are similar to those involved in interaction with touch enabled devices. Moreover, we review existing works in the Human-Computer Interaction literature and other fields that focused on the visual and manual modalities, as well as the concepts of Intelligent and Adaptive User Interfaces we consulted to construct the contexts and application objectives of our own research work.
- **Chapter 3 - Study Setup** : summarises the setup followed with the eye tracker devices we used for our data collections, as well as the techniques for filtering the on-screen hand gestures for selecting taps only for the first data collection.
- **Chapter 4 - Correlation between Gaze and Tap**: first core chapter of the thesis, it details the correlation between gaze and hand for taps, based on a data collection context containing Internet related tasks on a tablet. It also presents a pilot study to evaluate whether a commercial rack designed for eye tracking with small devices affects the naturalness of the interaction on the tablet.

- **Chapter 5 - Correlation between Gaze and Stationary Hand Event**: second core chapter of the thesis, it details the correlation between gaze and hand for stationary hand events (hovers and dwells), and also presents the role played by indecision in the interaction. A tablet is also used as a touch enabled device, but the context of the data collection is different. The data is collected while playing a Memory Game.

- **Chapter 6 - Correlation between Gaze and Hand in Motion**: third core chapter of the thesis, it details the correlation between gaze and hand when the hand is in motion (between taps and stationary hand events). The data collection context is the same as in the previous chapter.

- **Chapter 7 - Discussion**: offers a reflection on the thesis' core results and proposes future work guidelines that can benefit from this thesis' content.

- **Chapter 8 - Conclusion**: summarises the thesis' content and matches it with the initial research questions highlighted in the thesis' introduction.

2

Background

In the following chapter, we expose the background literature on which the matter of this thesis is supported. Since our work regards both gaze and hand modalities, Section 2.1 serves as an introductory presentation of the human visual system and eye tracking on one side, and then on the hand biomechanism and hand tracking on the other side. In Section 2.2, we further develop the separate literature of gaze and hand modalities by referencing works on the correlation between the both, in the fields of psychology and neuroscience (to understand the basic principles behind the correlation between gaze and hand), and Human-Computer Interaction (where the mouse instruments the indirect manual input instead of the hand). With Sections 2.4 and 2.3 we turn towards concrete examples found in literature to reflect on the applications of the correlation between gaze and hand in the area of Human-Computer Interaction and study contexts respectively, and relate them with our work.

2.1 Tracking Gaze, Tracking Hands

The following section instigates the fundamental knowledge of each modality covered in this thesis: the eye and the hand. For each, we present a physiological summary, followed by the key concepts and milestones of their tracking techniques, essential for the integration of these modalities in Human-Computer studies.

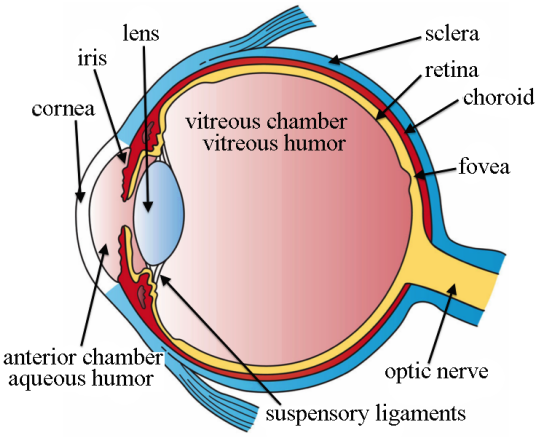


Figure 2.1: Structure of the human eye.

Source: Wikimedia Commons

2.1.1 Overview of the Eye Tracking

2.1.1.1 Description of the Eye and Visual System

The eyes and the visual system are a complex biological system in which the eyeball acts as a camera. Figure 2.1 illustrates the different elements that compose the eye.

The light reaches the eye's interior via the cornea and the lens, which play the same role as the lenses of a camera: the convergence of the lens directs the light inside the eye. For subjects with good vision, the cornea and the lens allow the light to be projected onto the bottom of the eyeball, the retina, to form an accurate image. Accommodation is the process by which the lens changes its vergence to adapt the projection of the image on the retina to the different distances between the eye and the point of gaze. The image formed on the retina is upside down, and the different receptive cells of the retina transmits a nervous signal to the brain via the optic nerves. The brain is then responsible to reconstruct the image and the visual interpretation from the nerve impulses it receives.

The eyes are paired in order to allow stereoscopic vision. Since the eyes are distant from each other by approximately 6 centimetres [205], each eye can see a scene from different angles, and therefore produces a slightly different image that is combined by the brain to create the perception of depth.

2.1.1.2 Vision Characteristics

The horizontal field of view obtained on a typical subject, when gaze is aligned with the head and centred, is about 200°. Of this angle, approximately 120° is covered by the binocular vision (at the centre of the field of view) [205]. Vertically, the total field of view is more limited (since the eyes are organised on a horizontal plane): Wandell in [205] reported a field of 135° (therefore, the field of view is wider in the vertical dimension in the binocular space).

Although the field of view is quite large, the human vision is not entirely efficient throughout that space. The retina is covered by different cells responsible to transduce the image (photons) to nerve impulses. The cones are the cells that detect colours. Three types of cones react to three different subsets of the light spectrum wavelength, corresponding to the red, green and blue colours. The cones are concentrated in the fovea, the part of the retina where the vision is the most acute, aligned with the visual axis. The other cells, the rods, are sensitive to dimmer light and are used by the system to interpret light intensity as well as movements. They are concentrated in the periphery of the retina and are therefore almost absent in the fovea.

The consequence of this distribution is that effectively, the eyes are only very efficient in a small part of the field of view, about 2°. Of course, the outer areas of the field of view are also taking part in the visual perception but do not allow such a detailed perception of the environment.

The eyes need to move in order to interpret a scene correctly. In eye tracking, three types of eye movements are usually reported or used: **fixations**, **saccades** and **smooth pursuit**. Smooth pursuit is the continuous eye movement triggered when a subject follows a moving element of the scene (for example, a bird flying). When a subject is looking a scene (without following a movement of the scene), the eyes still keep moving in order to analyse it. However, these movements differ from smooth pursuit: they are not continuous, and each "shoot" follows a ballistic trajectory. These shoots are called saccades. Between saccades, the eyes are keeping a *relatively* still position, in order to allow the visual input to be processed. These "pauses" are referred as fixations. Fixations can last milliseconds to seconds, depending on the activity. In eye tracking, fixations are often considered as an indicator of attention [94].

2.1.1.3 Tracking the Eyes

Earlier academic research in the ninetieth century experimented mechanical, intrusive and cumbersome devices to track the eye movements [94]. As the understanding of the eye physical characteristics and technology improved, better tracking devices have been designed through modern days. Some of the following works are listed by Holmqvist et al. [94], Wade and Tater [204] and Duchowski [56] as milestones in the eye tracking field.

Duchowski, in [56] categorises eye trackers in 4 categories: (1) EOG (electrooculography), (2) sceral[1] contact lens/search coil, (3) POG (photooculography) or VOG (videooculography), and (4) video-based combined pupil and corneal reflection.

A technical innovation of exploiting the light reflection by the cornea has first been used for eye tracking by Dodge and Cline in 1901 [53]. When light is emitted towards the eyes, the cornea and the lens reflect it in four points (the Purkinje reflections, of which the first reflection on the cornea generates the brightest *glint*). The interesting property of the reflections comes from their stability: if a light emission remains static, its reflections will be static as well on the eye, and by comparing the position of the pupil (for example) with the reflections, the direction of the eyes can be estimated. Their system projected a light beam onto the eyes, and photographed the concorded Purkinje reflections to evaluate the horizontal and vertical movements of the eyes. This system was the precursor of what is, to date, the most common eye tracker technology employed (type (4) as listed above). For instance, one of the most popular eye trackers on the market now, Tobii EyeX, is using a video recording of infrared beams' reflection onto the eye. The popularity of this tracking method certainly resides in its relative simplicity, its good performance without intrusion and its affordable price. Nevertheless, it is bound to known limitations, such as the ambient light exposure or the natural obstruction of the eye (droopy eye lids), and requires a calibration to generate a particular eye model whenever someone's eyes are being tracked (cf. Chapter 3 for details). The manufacturer of the remote eye trackers we used reports a latency of less than 35 ms (Tobii X2-60), whereas for information, a fixation typically lasts about 200 - 300 ms [94].

Other technologies brought by researchers paved the way of eye tracking. Lenses to record the electromagnetic induction provoked by the eyes were presented by Robinson in 1963 [166]. This technology was considered as a precise way of tracking the eyes but showed some inaccuracies during saccades and suffered from an obvious intrusion. As a

[1] On the white part of the eye globe, cf. Figure 2.1.

consequence, it is less favoured now. EOG, introduced by Schott [171] in 1922, has been widely used as it avoided the direct application of any measurement devices onto the eye (lenses), improved the comfort of the participants and provided an easy way to record the eye movements, based on the muscular electrical signal around the eyes area. The results of EOG were more reliable for horizontal eye movements, and were often subject to much noise, hence their seldom implication in current Human-Computer Interaction research works. The video tracking of the eyes has benefited from computational hardware and software improvements, and relied on the eye's feature evaluation without the requirement of a light beam emission [93, 153, 223]. A branch of eye tracking research still continues to work on video trackers to date, in order to improve detection algorithms and comply with hardware that can be easily found in regulars devices.

As we wanted to track gaze of the participants while they perform tasks on tablets, we checked how previous research works allowed eye tracking on such device. Drewes [54] mentioned that the first commercial eye tracker for tablets were used in his work, the ERICA[2] eye tracker (white[3] pupil method and a corneal reflection type), for his experimental study on a 12" (1024 × 768 pixels) tablet PC. Holland et al. [93] evaluated the usability of eye tracking on an unmodified tablet using a neural network and the tablet webcam, and reached an average spatial accuracy of 3.95°, and an average temporal resolution of 0.65 Hz with a 1024 × 768 pixels tablet. Likewise, Kunze et al. [119] introduced an open library for eye tracking on unmodified tablets using the webcam and a shape based image processing approach in the context of a reading activity. In order to avoid heavy development time ("discovering" a new device and its interface) and imprecision, we have tried to integrate a remote eye tracker technology often found in literature to our tablet: corneal reflection of infrared emission light. We favoured remote eye trackers over wearable eye trackers in our work because it was important for us to be able to track the users' gaze without obtrusive technologies: if gaze needs to be tracked in a public space, wearable eye trackers are not a realistic option to consider.

[2]Eyegaze Response Interface Computer Aid `http://www.eyetellect.com/` (last accessed Jan. 2020)
[3]When lit by infrared light, the image of the pupil obtained by infrared camera appears white.

2.1.2 Overview of Hand Tracking

2.1.2.1 Hand Biomechanism Involved in Aiming and Reaching

After the brain, the hand is probably the most crucial body element humankind benefited from nature: it allows fine dexterity (due to opposable thumb and fingers size ratio [80], necessary for a good control of tools - Aristotle [9] described the hand as "*an instrument that represents many instruments*"). The hand also takes part in the nonverbal expression that permits our body language [48]. When interacting with touch enabled devices, users may show two types of hand gestures: those that serve as input (tapping, panning, etc.) and those that indicate their cognitive activity (pauses, flipping of the hand, etc.). Thus, it is difficult to study the hand behaviour while aiming or reaching when subjects are put in a completely natural situation, as the hand movements may partly consist in body language expressions. This perhaps explains why research on hand movement often contains very basic and laboratory controlled tasks for pointing or reaching (along with the willingness to understand a precise and unbiased component of the human behaviour). Research in the human targeting mechanisms has not only be conducted to address human-centred issues, but was also the fundamental step in robotics to replicate the human behaviour.

Uno et al. [194] detailed the steps involved in one of the models that describes targeting movements: first the central nervous system determine one trajectory to the target that could be one out of an infinite possibilities, then the central nervous system translates the visual coordinates into body coordinates for the selected trajectory and finally the central nervous system triggers motor commands to the different elements of the body involved in aiming. According to Morasso [133], the human nervous system deals with the movements of aiming as the control of the hand trajectory rather than the control of the joints angular curves. His work limited the degrees of freedom of the body to the shoulder and elbow, but as a matter of fact, several other the body parts are also effectively involved in aiming (eyes, head, torso, wrist and fingers [88, 182, 198]), which rely on visual feedback and proprioception. Furthermore, the kinematic output of the limb movement is complex (control of the equilibrium of the hand under the actions of agonist/antagonist muscles involved [67]). Jagacinski and Flach [105] proposed the *bang-bang* model of the hand movement: the first *bang* corresponds to the acceleration needed to move the hand's mass from its current position towards the target; the second *bang* corresponds to the deceleration required to "slow down" the hand to reach its target.

In the context of a discrete target acquisition, optimal trajectory (in space and time) is achieved by the control of the switch between the two *bangs*. If the switch is made too early, the movement will fall short of the target. If the switch is made to late, the hand will overshooting the target.

2.1.2.2 Hand Tremor

In order to estimate the parameters implicated in the design of the algorithms we wrote for detection of the stationary hand events (cf. Chapter 5), we queried at how hand tremor was described in literature, since these algorithms are based on the natural hand behaviour Hand tremor is usually reported by the frequencies found in a spectral analysis. Literature provides several values: König et al. [116] reported oscillations in the 8-40 Hz range, Morrison and Keogh [134] reported 3 components: mainly 2-4 Hz and 8-12 Hz, but also at a smaller level, 18-25 Hz for the index finger. Morrison and Newell [135] indicated very similar frequencies (1-4 Hz and 8-12 Hz) with higher amplitude for the first range, they found a mean absolute acceleration of 0.06 m/s^2 for the index. In the same range, Strachan and Murray-Smith indicated that *action* tremors occur between 8 and 12 Hz; they proposed an Human-Computer Interaction application consisting in sensing the tremor pattern of an individual as a mobile device handling recognition technique [181]. Xia et al. [216] proposed a new measurement approach to evaluate tremor, and found (for the right hand) a tremor of 2.73 Hz. Finally, Tatinati et al. [186] modelled tremor and mentioned it was typically in the 6-20 Hz range, with an amplitude of 0.05-0.1 mm. Referring to the values reported by Xia et al. in [216] for healthy patients' right hand, we can estimate the station hand events to be comprised in a space of twice the amplitude (computed as the average of the X and Y amplitudes), or 9.96 mm, and a tremor velocity of 52.6 mm/s (computed as $4 \times amplitude \times frequency$).

2.1.2.3 Tracking the hand

The tracking of the hand started with mere observation (as would testify Ancient Greece's writings), just as the case of eye tracking, before technology could assist the lack of precision and information retention that humans can achieve with this basic methodology. However, the major difference with eye tracking is that the hands are protuberant and therefore easier to track without intrusion. The following history related to the tracking of the hand movements is a summary from Thurston's article [189] and Rautharay

and Agrawal's survey [164]. It seems that the body movements as such were not the key interest of research in the Greek or Egyptian civilisations: mainly, *what triggers* the movements was the real concern. Only a few steps in understanding the human body mechanisms were relevant until an anatomist, Galen, proposed a quite detailed description of the muscles by dissecting animals: he realised that the brain and the nerves played a role in the muscular activity (even if, at the time - second century A.D. - there were still misconceptions such as the humour theory he carried on from Antiquity, and religious taboos). The pinnacle of this anatomic research came with the Italian Renaissance: Da Vinci pointed out a link between mechanics and anatomy (bases of the kinesiology) and Borelli, considered today as the father of biomechanics, applied mathematical and geometrical principles to interpret the body movements.

Photography became the technology that drastically changed the way people studied movements. The infamous horse galloping pictures of Muybridge (1882) is a concrete example of how photography helped to analyse the different parts of the body movements while performing actions. In regards to human movements, he also produced a well-known extensive work that required a series of cameras, triggered sequentially, to decompose the walking movements and other human actions. Contemporary to Muybridge, Marey also used chronophotography with a movable camera to record people dressed in black, in front of a black background, with white markers on the body to analyse their movement.

From then on, two trends could be followed to track body movements: unobtrusive and intrusive. Unobtrusive works are solely based on image recording of the body. Computing technology brought another milestone in tracking since previous works could not provide instantaneous measurements, nor simple apparatus deployment. Tracking the limb relies on two approaches [164], detection based on the limb's appearance (skin colour model, silhouette model, motion based model, deformable gabarit model) or via applying a 3D model (3D textured volumetric, 3D geometric model, 3D skeleton Model). To date, the most commonly used commercial tracking devices are the Leap Motion, which computes a 3D skeleton of the hand based on two infrared cameras and light emission (depth detection); and the Microsoft Kinect which tracks the full body skeleton model based on a combination of ambient and infrared light sensors. Intrusive works handling the tracking of the upper limbs allows a "direct" measurement of the movement with the possibility to get good accuracy. Early devices included Karpovich's electrogoniometer (1959), which instantly reports the angular displacement of a joint and can be used as an exoskeleton. Biomechanics studies also use pantographs (for the hand trajectory measurement) and

potentiometers (for the joint angular measurement) in targeting activities. Tracking the hand is also often done with gloves or rings that contained markers or embedded accelerometers.

Along with the use of computing devices, screens became a tool for tracking the hand, in a relative small space portion above the device. Initially digitizers used stylus to reproduce the 2D hand movement given to the stylus. In *The History of Visual Magic in Computers* [155] presented the milestones of digitizers' development. The first device (Telautograph) considered to reproduce a hand movement has been patented in 1888 by its inventor, Elisha Gray. This technology allowed the hand movement to be "tracked" via the stylus in in 2D but relied on the pen movement solely, contrary to tablet digitizers for which the surface is a part of the tracker and the stylus a complementary tool working with the surface. Therefore, the first tablet digitizer to be developed as a tracking surface with a stylus is the Stylator (1957), followed by the RAND tablet (1964) which gained more popularity. The RAND tablet is descrbied by Peddie as a tool that "*employed a grid of wires under the surface of the pad that encoded horizontal and vertical coordinates in a small magnetic signal. The stylus would receive the magnetic signal, which could then be decoded back as coordinate information.*". Based on the same technology principles, the BitPad was commercially successful in the late 70s and early 80s thanks to the introduction of cheaper and faster computing power. Branded for Apple (as *Apple Graphic Tablet*), they allowed the detection of the stylus tip in a small range *above* the surface thanks to magnetostriction. [4]. Nevertheless, such devices remained detached from the computer itself until the introduction of GridPad 1910 in 1989 (according to Peddie, it can be considered as "*the first commercially available tablet*"). GridPad allowed a basic interaction (form filling) thanks to the stylus tracking over a monochromatic touchscreen display. On the most iconic tablet digitizer was introduced by Microsoft in 2001 as the *Tablet PC*, that still included the use of a stylus for pointing while the finger could also be directly used. Two main touchscreen technologies can be found: capacitive and resisitve. Capacitive technology for finger interaction with a tablet was presented by Johnson [110] in 1965. A capacitive screen detects the finger thanks to its electric conductivity and dielectric difference with air: the finger acts as a capacitor and distort the electrostatic field generated by the screen. The first resistive screen was presented by Hurst and Colwell [98] in 1975. The principle behind resistive screen is the creation of a voltage divider by contact from one layer of the screen (the touched

[4]Wikipedia article on Graphics tablet `https://en.wikipedia.org/wiki/Graphics_tablet` (last accessed Jan. 2020)

one) onto the other underneath that is normally separated by a gap when no touch is performed. The stylus, in current technology, has been mainly discarded for every day use since the interfaces and the capacitive screen technologies allowed enough precision when the finger is used for touching. The finger or a stylus can also be tracked in the Z-axis with capacitive technology in a small range above the surface (due to the distortion the hand or the stylus creates in the electrostatic field of the digitizer). Microsoft proposes a test tool to measure the hover range of a stylus[5] with an acceptance threshold of 5 mm. In Mobile Word Congress 2013, STMicroelectronics presented a touchscreen allowing hovering at 2 inches above the surface.[6] Du et al. [55] proposed a touch sensing circuit for mobile devices reaching a hover range of 11 cm with a centimetre resolution. In [89], Hinckey et al. used a prototype mobile handheld device (based on the Fogale Sensation) which permitted a hover sensing of 35 mm above the screen surface.

Commercial tablets, such as Microsoft Surface, are good tools to track the hand's taps on the surface. For the work related in this thesis, we also wanted to track the hand *above* the surface. To avoid obstruction (both from hardware and in the user experience i.e. with calibration) we favoured Leap Motion to track the hands in the air for its reliability, ease of use and remoteness.

2.2 Correlation between Gaze and Hand

This section is dedicated to the principles of the correlation between gaze and hand established by research on the human behaviour. We organise the presentation of these works as follow. First, we detail the general concepts of the correlation between gaze and hand in aiming and reaching, studied in psychology and neuroscience studies, in order to apprehend the basic operation of the correlation in simple context detached from Human-Computer Interaction, and also to understand what are the key features to investigate when working with it. Secondly, we outline the research activities that employed the correlation between gaze and hand in Human-Computer Interaction, to appreciate how it was approached and why.

[5] Microsoft Hardware Dev Center (Hover Range test) `https://docs.microsoft.com/en-us/windows-hardware/design/component-guidelines/hover-range` (last accessed Jan. 2020)

[6] `https://www.cnet.com/news/touchless-touch-screen-gives-you-control-without-contact-video/` (last accessed Jan. 2020)

2.2.1 Gaze and Hand Coordination in Aiming or Reaching

Aiming at and reaching for an object involves both the hand and the eyes. Even if they are trivial actions of the human activities, their "*functional organisation [...] is not yet fully understood*" according to Vercher et al. [198]. An early study on the control system involved in the generation of hand movements to reach objects has been done in 1899 by Woodworth [214]. Woodworth proposed "the two components" model in which he considered the hand movements were controlled by two components: a central part and a feedback part [61, 214], and the core of his work was to understand the relationship between accuracy and speed of the controlled goal targeting limb movements. In relation with eye-hand coordination, Woodworth found that visual feedback helps to increase the accuracy of the movements, but only when speed is low enough (movements approximating 450 ms). In their review, Elliott et al. [61] investigated whether this model was still available a century later, and mentioned key research works that supported or contradicted the two components model. They concluded that Woodworth's framework was still valid but required to be completed by finer models. Some of the works they reviewed are listed in the following paragraph.

Intuitively, we would assume the eyes first acquire the target before the hand moves. If the eyes indeed reach the target prior to the hand [1, 16, 17, 198], the coordination between hand and gaze is more subtle and concurrent. Vercher et al. [198], in their study focused on the correlation of the more extended system *eye-head-hand*, summarised three steps in the pointing process: "*(1) location of the object with respect to the body, involving coding target position with respect to the fovea, the eyeball with respect to the head and the head with respect to the body; (2) knowledge of forelimb position by means of proprioceptive and/or visual afference; and (3) coordination of eye, head and arm movements leading to gaze and arm shifts towards the target.*". Literature indicates that, however, the hand is not a *slave* to the eyes [17] but that the eyes are the precursor of the information needed for the hand to reach a target with better accuracy. In other words, they work together. Researchers have tried to understand how this relationship was organised to model the human visuomotor process involved in what seems a very trivial task. According to literature, in target acquisition tests, the gaze leads the hand in a reaching task by 60 to 100 ms [16]. Fischer and Rogal, in [62] reported different reaction times explained by the gap and overlap paradigms: if the reference light was turned off before the target light were shown (gap paradigm), the gaze saccadic reaction time was

shorter (120 ms), whereas it increased to 200 ms when the reference light was not switched off during the appearance of the target light (overlap paradigm). Keele and Posner [113] conducted a study to evaluate the visual feedback duration in rapid movements. Using a comparison between two target acquisitions where the actions were performed in a fully visible condition and when it was done in the darkness, they concluded that the processing of the visual feedback takes between 190 and 260 ms. Carlton, in [30], studied the contribution of the visual information. He indicated that the common model, at the time of writing, was that visual information is needed to correct the errors of the hand, and thus, the eyes actually need to monitor the hand itself. However, he also explained that other researchers (as Stubbs [182]) have mentioned that aiming accuracy should not rely on the visual monitoring of the distance between the target and the hand, since the hand position is already known by the motor system from proprioception. His work concluded that the proprioception does not prevent the users from visually monitoring their hand (and target) for better accuracy in aiming. The impact of the stimuli and potential distractors on the hand movement were analysed in [31, 175], and concluded that they infer the hand trajectory. Abrams et al. [1] have demonstrated that distance evaluation by the eyes depends on their position, and how gaze accompanies rapid limb movement (wrist rotation for pointing): they found that moments before, quite simultaneously with the hand movement, the eyes would perform a saccadic movement towards the target, and tend to undershoot the target, requiring another small saccade to adjust the destination goal. The limb movement is also proven to be undershooting [61, 176], and this behaviour is explained by researchers as a result of time and energy economy by the human system. We can suppose the same explanation is valid for gaze movements, as mentioned by Hansen et al. in [87].

Despite bringing valuable insights in the understanding of the gaze/hand correlation during hand movements, research often lacks example of studies that reflect a completely natural activity. For example, measuring how the hand performs a targeting action with a stylus as apparatus may be questionable when interpreting results that describe direct manual reaching (because the stylus may add another control factor in the movement, whereas the hand physical characteristics is already known by the control system thanks to proprioception). This thesis explores the correlation between gaze and hand using a tablet as a touch enabled device, which interaction from the users simply requires the most natural way of targeting: direct hand input.

2.2.2 Mouse and Gaze

We primarily interact with computers using our hands (as input) and our eyes (as output). Before the generalisation of touch enabled devices, understanding the correlation between gaze and hand in the field of Human-Computer Interaction relied on the mouse, which acts an indirect manual input. Early work on gaze and mouse input correlation in Human-Computer Interaction can be found in [174] (where Smith et al. studied the correlation in target selection tasks and found out several patterns) and in [33] (where Chen et al. found correlation patterns applied to web browsing tasks and an average correlation of 0.58). In an exploratory experiment, Cooke [47] showed that mouse and eyes are correlated 69 % of the time the mouse was on screen for different search tasks on a set of webpages. Huang et al. [97] tailored a finer experiment to understand when gaze and mouse are aligned in search tasks, and proposed a model of gaze prediction. Buscher et al. [27] studied eye and mouse correlation on Search Engine Results Pages (SERP)'s advertisement content.

When interacting with a desktop computer, the hand does not manipulate the mouse continuously. Therefore, strictly from the device's point of view that only receives mouse signals from a user, the interaction remains sometimes idle. Rodden et al. [167] described the coordination patterns on web search and noted that users often leave the mouse to unmeaningful regions (labelled "nowhere" by [33]). Therefore, questioning the impact of the task's nature on the correlation is interesting. Bieg et al. [15] studied the correlation between mouse and gaze on abstract search and selection tasks and showed that initial knowledge of the target location influences the eyes targeting, and that users perform search and pointer movements simultaneously when the tasks require visual search of a target item before selection. They also confirmed the findings of [174] that eyes reach the targeted item prior to the pointer, and they extended these findings by noticing that the eyes tend to fixate the targets rather late when the approximate target location is known. Liebling and Dumais [123] explored the correlation of eye and mouse in everyday computer work tasks, which is another form a natural Human-Computer Interaction study. They confirmed the fact that the eyes lead the mouse but nuanced this paradigm indicating that it occurs only two thirds of the time, as "*[this] depends on the type of target and the familiarity with the application*". Çöltekin et al. [46] investigated the correlation of mouse movements and gaze with visual search tasks on geographic displays in order to estimate gaze position from mouse movements.

We contribute to the understanding of the correlation between user's tap and gaze by studying *direct* touch input instead of the mouse, the traditional indirect manual input of the works listed in this section. Besides, these references provided examples for the data analysis of the correlation between gaze and hand in a Human-Computer Interaction context.

2.2.3 Correlation between Gaze and Hand Movements

In the field of Human-Computer Interaction, a lot of studies scrutinised the correlation between gaze and mouse (i.e. [15, 33, 34, 46, 47, 77, 82, 97, 101, 123, 141, 142, 167]), claiming that the mouse is an indirect manual input. However, when just paying attention to the motion part of the hand/mouse, these studies are not a good representation of the correlation between gaze and hand. The mouse may be at an unknown position that the eyes need to capture (roughly) again, whereas the hand position in space is always unconsciously known by the users thanks to proprioception. Studies that pinpoint the hand and gaze correlation during hand movement are, because of the naturalness of the interaction, therefore found extensively in other fields of research (psychology or neuroscience), in order to understand and model the human behaviour. Comprehending how gaze and hand correlation during hand movements can be studied from two complementary angles: in the temporal or in the spatial dimensions (Neggers and Bekkering wrote in [143] "*The central nervous system (CNS) apparently enforces a co-alignment of the ocular and manual motor systems in space and time.*"). Spatially, Fisk and Goodale [64] showed that the hand approaches straight line paths at an inconstant speed between two reaches. They also demonstrated that the latency between the gaze and the hand to reach a target depends on directions: the distance was smaller when the hand reaching the target was the same side as the gaze direction, longer when the opposite hand was used. Neggers and Bekkering [143] found that gaze stays locked on the target during pointing even when a new target appears during the movement. Keele and Posner [113] highlighted the problematic protocol of the reference work from Woodworth [214] in which the hand was doing back and forth movements. They suggested that the nature of reversal movements brought some inconsistency in the analysis of the hand movements as the forces deployed to counterbalance the inertia of the hand required extra time (and energy [81]). Common sense would suggest that the temporal organisation of the eye and hand while targeting necessarily means the eyes leads the action. In natural activity contexts, Land [121] confirmed that the eyes led in the organisation of an action but

there was no focus on the hand behaviour as such. Keele and Posner [113] evaluated the time required for the central nervous system to operate after the visual acquisition to last between 190 ms and 260 ms. Indeed, the role of the eyes relative to the hands during hand targeting is subject to a debate: [16, 113, 143] supported the later statement (meaning that the eyes "command" the hand in the control system), whereas other researchers [17, 18, 198] rather considered that the hand and the eye are *simultaneously* and *collaboratively* participating to the reaching a target. There is however no doubt that gaze reach eventually the target faster than the hand [1]. In their review work, Elliott et al. [61] reported that research studies showed how gaze contributed to maintaining the hand's accuracy by corrective movements controlled by the CNS.

2.3 Study Contexts

People commonly interact with touch enabled devices: mobile phones and/or tablets in the private sphere, and kiosks at the public areas. Therefore, it would seem illogical to study the correlation between gaze and hands based on an unnatural and unusual interaction, such as performing an abstract task. During the preparation of our data collections related to gaze and hand, we have devised contexts that are also found in literature, since examples of natural and common activities are numerous in other works which studied, among other interests, the correlation between gaze and mouse. The first data collection context, described in detail in Chapter 4, is related to Internet based activities and the first part of this section lists the research works that inspired us for constructing our own context. The second data collection context was influenced by the need of generating enough cognitive load from our participants, as explained in Chapters 5 and 6, and resulted in the deployment of a Memory Game. The second part of this section summarises some of the works we queried to find what decision making activities can be designed for research purposes.

2.3.1 Internet Based Tasks in Research

Studying how users interact with Internet seems very coarse, since many different types of activities can be performed by users. In [42, 51, 112, 120, 211], no specific task was designed because the data was collected "in the wild". These approaches implicated the users to run a client-side tracker, which in most cases did not interfere with the users' browsing habits. For instance, de Santana and Baranauskas [51] made an evaluation of

the different client-side tracking tools and introduced another one, WELFIT (Web Event Logger and Flow Identification Tool). Lagun et al. [120] collected data from the EMU[7] browser plugin installed on a library's computers to track their mouse cursor movement details. Claypool et al. [42] used both an unobtrusive tracking and intrusive webpage evaluation tools, as did Kellar et al. [115] (although the intrusive tool was designed to categorise the webpages at the time of reaching them instead of a popup window appearing after the page was quitted to evaluate the page content). White and Morris [211] used a client-side solution as well to collect data from free browsing activities. In [28, 108], the data was directly retrieved from the servers of two different search engines for query and search behaviour analysis. Arroyo et al. [10] introduced MouseTrack, a tool deployed on a webpage to retrieve mouse activity. Conducting "in the wild" data collection certainly gives the advantage of obtaining a large scale population, but limits analysis tools to the browser only in the case of client-side studies, or limits the nature of the study tasks in the case of server-side studies.

In fact, literature about user behaviour studies related to Internet based activities focuses very often on search tasks using SERP (Search Engine Responses Page). Search tasks performed by users can mainly be categorised as **informational**, **navigational** or **transactional**, as described by Rose and Levinson [168] and Broder [21]. Both studies showed that navigational searches are less common than transactional searches, and even less than information searches. However, as we describe below, studies involving SERP often combine a set of navigational and/or informational tasks only, when transactional tasks are preferably done on websites other than search engine websites.

In most cases of studies treating Internet search activity and SERP, participants were asked to accomplish search tasks by entering an initial query with a predefined search engine. The tasks usually consisted in answering *informational* questions [13, 69, 72, 83, 95, 142, 146, 167], which may have been combined with *navigational* questions [50, 57, 82, 97]. According to Buscher et al. [27], the queries were initially provided to the participants "*in order to make the initial SERP comparable across [them]*". Giving initial queries were also chosen in [50, 57, 82, 83, 97, 142, 167]. For the same reason, the search engine may have been imposed to the participants, regardless being a commercial search engine as in [13, 52, 69, 72, 82, 83, 97, 146] or a home designed one in [27, 57, 142, 167]. Some studies, as [27, 50, 57, 82], preferred using a cached SERP associated to the initial query in order to avoid inconsistency between the participants' experiments. To avoid

[7]The Emory System for Managing User Behavior data.

personalisation effect, the browser's cache, history and cookies were cleared in [97].

Search tasks do not always involve SERP. For example, Cooke [47] proposed a set of navigational search tasks on a governmental website to collect data. Similarly, Atterer et al. [12] asked the participants to find an entry to a particular point in Wikipedia's FAQ. Shrestha [172] analysed mobile web browsing usability relying on navigational tasks for which the target websites were indicated (i.e. BBC, Yahoo, National Rail). Nakamichi et al. [140] asked the participants to perform the same informational search on five different websites they had selected beforehand, while Jones et al. [112] or Griffiths and Chen [77] asked participants to look for different informational retrievals in the same website or portal. Kekäläinen et al. [114] gave the participants several informational tasks to be retrieved in a set of webpages; this method can also be found in [26]. In a holistic study, Wang et al. [206] gave the participants the informational tasks, but let them interact at will to complete the tasks, meaning their use of a SERP was uncertain. In those studies, except in [206], participants were always directed to predefined websites, even if [172] stated that participants were free to complete the tasks using the websites of their choice rather than the suggested ones, it does not seem they did so.

Broder [21] describes *transactional* search tasks as tasks where the search target is a website offering a transaction, such as an online shopping website, or the retrieval of a resource, such as a map or a file. Few studies focus on transactional search (in [95] the searched target was a compatible file for a software). Indeed, in studies involving transactional tasks, it is far more common to directly target a website in which "actions" can be performed, rather than using a search engine as an intermediate. Another classic Internet activity found in studies is shopping: participants were asked to use an online shopping website to find a product matching some criteria in [7, 11, 25, 41, 99] or to simulate the purchase of a product of their choice in [185]. Chudá and Krátky [41] proposed the setup of a "gamified" version of an e-shop to invite the users to interact with the website intensively. Using transportation websites in studies can also bring a lot of interaction from the participants. Despite not describing their tasks in details, Burzacca and Paternò [24] employed an airline website to study mobile web applications, Ehmke and Wilson [59] made use of a train tickets sales website. Online mailing activities are also requiring high interaction from the participants. Shrestha [172], or Atterer and Schmidt [11], included an email writing task in their studies, Iqbal and Bailey [99] devised a mail sorting task in different mail folders with drag & drop interaction to study object manipulation. Other kinds of study contexts call for a great amount of interaction from

the participants as well, such as manipulating a company's sales web-application [75], entering events in an online calendar [11, 12], evaluating a website's traffic figures in *Google Analytics* [49], playing a quiz on an e-learning platform [49], or following a path on a map to reach a target by using links [40].

Reading behaviour is another aspect of the research covering Internet activities, where participants were asked to browse the indicated websites with further feedback tasks [8, 25, 141, 151] or not [33, 146].

Instead of focusing on a particular type of activity, we gather the most common ones in our data collection context to cover a wider range of Internet based activities that also reflects the tablet's use in real life.

2.3.2 Decision Making Study Activities

The decision making process during target selection on computing devices involves both the human system (cognition), and the context of the stimuli on the machine [221]. So decision making activities are found in several fields, depending on which part of the process the research is focusing on. Psychology, economic and medical research focus on the human system, such as the arrangement of a decision making process [29] or the impact of ADHD on decision making [43] based on the results of recognised neuropsychological tests. Eye tracking is also sometimes used to assess the level of indecisiveness of individuals, for example in [125], where Lufimpu-Luviya et al. present different alternatives in a given context was presented on a screen to the participants, their choices were recorded by the experimenter from their oral answers; or Patalano et. al [154], who used a similar procedure, but the participant's decision was done via a game controller button.

Centred on computing, decision making studies either evaluate ways to assess or avoid indecisiveness, or explain what triggers it. For instance, detecting frustration was presented in [3] using sensors to detect typical hand features while participants play a game (in which glitches mean to provoke their frustration) . In his PhD thesis [2], Anh describes the experiment environments used for his work on the role played by emotions in decision making. One of them, involves, for the decision making part, a betting application participants played on a desktop computer. To monitor their response, a wristband skin conductance sensor was worn by the participants on their non-dominant hand. In a study of the effect of interruptions on decision making, Speier et al. [177] devised a

set of tasks (simple and complex) to be performed on a desktop computer. Simple tasks consisted in answering questions in view of preparing scheduling workload on several machines, while complex tasks involved accomplishing mutually related activities of facility location and planning aggregation. No body sensor was used in their experiments, they measured performance from the decision accuracy and time. Gonzalez [76] investigated the role of animation in user interfaces of two applications (for finding an accommodation rent, for playing a scenarios game based on a physical phenomenon) in regards to decision making. Ho and Tam [92] studied the effects of web personalisation on the decision making with first asking their participants to make selection of a sound media file from a web service.

Our work spans through both worlds as we describe the gaze and hand behaviour that characterises decision making on a tablet. From literature, we notice that the decision making context application varies a lot depending on the measurement made in the research work. Since we only need the participants to make decisions without measuring anything else but their gaze and hand movements, we selected an application that keeps the participants entertained while requiring them to make decisions: a Memory Game.

2.4 Human-Computer Interaction Applications

In this last section, we review the research works, from literature related to the field of Human-Computer Interaction, that illustrate *applications* from either using gaze and hand inputs (as two independent but complementary modalities, or taking advantage of the correlation between gaze and hand), or integrating above the air manual input. From this section, we want to justify the important roles gaze/hand as modalities and the hands in the volume above the interactive surface can play when interacting with computing devices - which corroborate the overall aim of this thesis. Therefore, this section is divided in two parts comprising each of these three centres of interest, that are directly connected with our research topic: applications from gaze and hand modalities, and applications from hand input above the interaction surface. Despite our thesis being organised around three hand events, the background review in this section only covers two: stationary hand events and hand motion. Contrary to the hand events above the air, tapping is by nature a very simple, commonly employed and direct interaction mean. In the third part, dealing with above the air manual input applications, we include a particular application: the detection of hesitation. One of the objectives of the

thesis is to understand how the correlation between gaze and hand can inform the user cognitive state, in particular hesitation. We selected works studying how the hands can indicate hesitation to find out if, as we suggest in the thesis, hand stationary events may be considered as an indicator of hesitation. The outcome of understanding the user's cognitive state, in Human-Computer Interaction, leads to the deployment of adaptive interfaces. The systems react the the users (intelligent HCI). Therefore, finally, we review the key concepts and state of the art experiments found in research related to adaptive interfaces.

2.4.1 Applications from Gaze and Hand Modalities

2.4.1.1 Using Gaze and Hand Modalities Independently

It is not a surprise that Human-Computer Interaction experts took interest in using gaze as an input. Jacob [103] made a detailed study of gaze use for Human-Computer Interaction which served as a reference for eye tracking, he pinpointed the "Midas Touch" problem that gaze interaction brings. One of the first major work reference related to gaze interaction has been provided by Zhai et al. [222], where they clearly explained what gaze interaction can bring: hands substitution, increased speed and health issues prevention by avoiding manual contact with devices in public spaces. Their work is often considered as one of the first concrete application examples of gaze interaction: they presented MAGIC, a pointing tool taking gaze as an active input.

Recent studies involved the hand directly: researchers in Human-Computer Interaction also acknowledged how much potential can gaze and hand bring together to the interaction with computing devices. Yoo et al. [220] designed a system retrieving hand and head movements (computed with depth and colour cameras, gaze was approximated from the head posture) to interact with a wall display. Gaze was used to target the centre of interest while hands performed actions such as zoom or panning (browsing) triggered by a push/pull horizontal movement. Similar work was presented by Slambekova et al [173]: eye tracking was estimated from an eye tracker and the hands gesture or movement from a Microsoft Kinect; and Chatterjee et al. [32]: using a remote eye tracker and a Leap Motion to explore continuous manipulation (`Gaze+Gesture`, where hand movement was used for moving a cursor). Velloso et al. [197] conducted a comparison between the different selection methods for object manipulation in a 3D space on a laptop screen: with gaze, with a hand 2D raycasting technique and with a 3D virtual hand technique. They

found that gaze and hand interaction performed better and in a more natural manner that the two other techniques. Same conclusions were shared by Zhang et al. [224].

Tablets' widespread and ease of use have also led researchers to find ways of estimating the point of gaze onto the tablet' screen without specific other tracking devices. This is the device we selected to represent touch enabled devices in our study. Despite no work has particularly focused on the correlation between gaze and hand on tablets, interest in using both input methods has been subject to several research projects. For instance, Turner et al. [192, 191] have designed a cross-device content transfer method using an eye tracker, from a shared monitor display to a tablet (or a laptop). In their setup, content acquisition and transfer were done by a combination between gaze and hand inputs. Pfeuffer and Gellersen [161] explored the potential of combined gaze and hand inputs for new interactions techniques on tablets, transforming the touch input as an indirect medium. Takahira et al. [183] designed a system using on a camera (viewing from the back of the user), that modelled gaze estimation based on several parameters, including the device's handling and eyeball position's estimation.

In these applications, however, the actions of the eyes and hands seemed complementary but somehow independent: no *combined* data served the systems to bring new information.

2.4.1.2 Using the Correlation between Gaze and Hand Modalities

In early works studying the correlation between gaze and hand, researchers focused on the mouse as the indirect manual input. Guo and Agichtein [82] acknowledged previous work on eye-mouse correlation, and they also proposed to automatically infer gaze position from users' mouse movements in the context of web searching activities. They obtained an accuracy of 77 % within 100 pixels. Navalpakkam et al. [142] studied nonlinear webpages layout (the content of the pages were not made of a single type of elements on top of each other in the manner SERP was typically presented then, instead, the pages were made of diverse elements organised in both dimensions of the pages) and they showed that both gaze and mouse are sensitive to two characteristics of the page elements: their position and their relevancy to the user's task. Additionally, they also achieved a gaze prediction based on mouse activity with 67 % accuracy, with an error up to one page element.

To our knowledge, gaze and hands *together as one unit* described by their correlation with one another has not been used for interaction involving the hand as a direct input.

2.4.2 Applications from Manual Input above the Interactive Surface

In the following, we summarise the works only involving the hand has an input *above* the surface of the interactive system.

Marquardt et al. described the possible interaction techniques combining gesture and touch with digital surfaces in [128], and cited hovering as one of these interaction means, for instance allowing feedback of possible actions.

Wacharamanotham et al. explored the finger interaction above the desktop devices in [203]: they first studied the best thickness of the volume above the desktop surface for proper interaction, and recommended a height depending on the nature of the hand movements (2 cm for short movements, 4 cm for long movements). Choi et al. targeted mid-air interaction on laptops: in [39] they designed a touchpad able to detect the hands above it, and they suggested the use of hand hover as an activity recognition (in their case, typing), but they did not detail the hover detection method.

In [213], Wilson and Benko proposed an interactive space in which the volumes between surfaces was exploited. Among different gestures, they used hand dwelling to validate a vertical menu selection. Even if they mentioned they processed images from depth cameras, the evaluation of a dwell was not detailed. Active pointing is closely related to "unaware" hovering; in [139], Müller-Tomfelde detailed the temporal steps of pointing, and reported dwell times from 300 ms to 1 s. However, dwell time comprises a reaction time from stimuli that is not existing in passive unaware hovering. Likewise, Colombo et al. designed an unobtrusive system (PointAt) to control the information displayed on a room's interactive walls using hand pointing [45]. They employed video processing and triangulation to detect the hand direction (and pointing), as well as thresholds to ensure the hand was actually engaged in pointing: a temporal threshold indicating the minimum duration for which the hand stayed in a limited portion of the screen (spatial threshold). Their work was not focused on the evaluation of these thresholds.

Hover is often used for interaction with mobile phones. For instance, with Air+Touch [35], Chen et al. explored different gestures in the air and hover was used to control the pages scroll speed. Nevertheless, again, the hover detection as such wasn't detailed. Hinckley et al. [89] used pre-touch sensing on a mobile phone based on its touch screen

and edges detection. Despite not using the stationary events of the hand, some of the techniques they proposed, such as detecting the finger's *approach* to place a menu at the right position, could be also triggered by a dwell of the hand (they employed *hover* as an equivalent to mid-air interaction, as do Ostberg and Matic [148] who designed a mid-air pointing and selecting technique on mobile phones).

Grossman et al. developed a device to interact with thanks to multi-fingers gestures [79]. A spherical display tracked the hands with markers in order to manipulate objects in a 3D environment. Hand motion, combined with gesture for object selection, was used here for moving the objects. Wilson and Benko proposed another prototype of 3D interaction system, `LightSpace` [213], comprising two interaction surfaces (vertical and horizontal) in a cubic structure (smart room), where the volume within was also an interactive space. Hand movements were used to move objects around (on, between and above the surfaces). Cheung et al. [37] discussed the interest of finger hovering on a tactile display to improve interaction, in the same manner mouse hovering brings extra information on a desktop computer, even if in their speculations, hand motion was limited to analyse the finger's approach. Han and Park [84] designed a hover technique in line with Cheung et al.'s aforementioned speculations. They proposed two zoom techniques relying or partly relying on hover distance to a tabletop screen (estimated by computer vision). Wacharamanotham et al. [203] explored the mid-air interaction above devices (keyboard) for standard desktop configurations. They searched in which volume hover was reliable to be used without ambiguity, and tailored an in the air clicking technique above the keyboard. Choi et al. targeted mid-air interaction on laptops: in [39] they designed a touchpad able to detect the hands above it, and they suggested the use of hands hover as an activity recognition (in their case, typing). Xia et al. [215] recorded hand movements with a ring sensor to evaluate the trajectory during tapping on a large tactile display (with controlled starts and ends) in order to model this trajectory and then build a touch predictor to reduce the latency between tap and system response. Similarly, Onishi and Shiroshima [147] collected the projected trajectories of the hand while performing taps on a smartphone's touch screen (hover detection built in) to prefetch the data to the then estimated touch point. With the willingness to develop new interaction techniques in mind, Marquart et al. [128] proposed a range of gesture to work with a (large) touch enabled screen, exploiting the space above the surface to manipulate objects or trigger actions from the interface. Hinckley et al. [89] explored the hand proximity sensors to modify a phone's interface at finger approach; while Chen et al.

[35] developed `Air+Touch`, a set of new techniques for mobile phone interaction using a depth camera to record in the air gestures, and suggested application examples on maps, document reader, photo gallery and drawing. Ostberg and Matic [148] facilitated the performance of taps on small targets by showing continuous finger's projection feedback using the hover sensing of a mobile phone screen (`Hover Cursor`).

This short review shows the potential applications of retrieving hand motion. We are taking a step further by also retrieving gaze and thinking how the combination of these two inputs can serve Human-Computer Interaction in a smarter way.

2.4.3 Hesitation Detection

Observations associated with Fitts' law related studies informed that during continuous target selection, the hand realises "dwell times" between two consecutive movements [66]. Meyer et al. [130] highlighted the role of hesitation in hand dwell time. Interpreting human hesitation has been studied in Human-Robot interaction. Moon et al. [132] investigated hesitation characteristics in a targeting conflict inter-human collaborative activity, to later implement this behaviour to a robot. They modelled one type of hesitation (retract) based on acceleration to evaluate nonverbal communication with robots. Their study, however, did not focus on more than one target, and they showed that their model could not work with the "pause" type of hesitation.

Nevertheless, their work is also used in Human-Computer Interaction, as explained by Vodlan et al. [201], who made a clear explanation on how Social Signals Processing can be used for Intelligent Human-Computer Interaction [58] (HCI^2 [152]), and in particular how gestures can indicate human hesitation to a machine. They classified the subjects' activity into two states, {hesitation | no hesitation}, based on their observation, and proposed a logistic regression model relying on the most significant observed features [200].

Time indicators between stimulation and response can show hesitation and therefore be used by machines too [136, 201]. Our work adds-up with this research trend by proposing a method to evaluate hesitation based on different input channels (gaze and hand).

2.4.4 Intelligent and Adaptive User Interfaces

Human-computer Interaction encompasses the technologies, practises and understandings that allow humans to work with machines, but also machines to work with humans (human-centered interaction [149] or Intelligent Human-Computer Interaction [152]). Jameson [106] describes a user-adaptive system (equivalent to "adaptive interfaces", "personalisation" or other terms used in literature) as a system that "*makes use of some type of information about the current individual user [...], can be defined as an interactive system that adapts its behavior to individual users on the basis of processes of user model acquisition and application that involve some form of learning, inference, or decision making*". Thus, the adaptation may result in different outcomes: sometimes clearly visible (such as the content adaptation when the content depends on the user's activity, choices and habits i.e. commercial website's item suggestions) and sometimes not visible (such as the system adaptation where the processes are adapted to the user to guarantee an optimum result commonly expected by the system i.e. phone typing auto-correction). Benyon, in [14], gave the guidelines to analyse the usability and to design systems appropriately to "*build intelligence into the system*" and reports the four adaption levels established by Browne et al. [22]: *simple* ("*use a 'hard-wired' stimulus-response mechanism*", *self-regulating* ("*monitor the effects of the adaptation on the subsequent interaction and evaluate this through trial and error*"), *self-mediating* ("*monitor the effects on a model of the interaction*") and *self-modifying* ("*capable of changing* [their] *representations*", the models can be adapted). The adaptation of a system may take different aspects: *taking over parts of routine tasks*, *adapting the interface*, *helping with system use*, *mediating interaction with the real world* and *controlling a dialogue* according to Jameson [106]. In the following background review, we refer to the major research works focusing on the particular functionality of adaptive interfaces.

Early personalisation interfaces have been experimented by Chesnais et al. [36] and Höök [96]. Chesnais et al. introduced *Fishwrap*, a personalised electronic newspaper targeting freshmen at MIT. *Fishwrap* delivered content based on the user's personal information (place of origin, affiliation with MIT and interests) and interaction (position of the consulted articles within the page), and had been well received by the readers despite concerns dealing with privacy. Höök presented *PUSH*, an adaptive hypermedia system, which content (information the users wanted to retrieved) was tailored based on the user's activity (clicking, StretchText[8] actions). The evaluation of *PUSH* revealed

[8]StretchText is similar to hypertexts but expand in-context, cf. `https://en.wikipedia.org/wiki/`

that users required fewer actions compared with the non-adaptive equivalent system and it was prefered.

Bohnenberger et al. [19] investigated the usability of a location-aware shopping assistant application on a PDA. Their application indicated the shopping areas of interest (adaptation of the content) for a user based on her location in space, but also on the interest showed in specific products and the purchases during previous shopping sessions. They showed the application performed better (faster shopping and preference) than a paper map.

Content adaptation is often deployed on commercial Internet websites to provide the buyer a fast access to their preferred or likely to be preferred products, ease the purchase process and run marketing strategies. Alpert et al. [6] conducted an evaluation of a user-adaptive online commercial website (prototype) from the users point of view, and found out that users did not always respond in favour of system intrusion and personalisation, and that they preferred "*[having] full and explicit control of data and interaction*" as well as clearly understand the way personalisation was in place on the system.

Content adaptation is not solely based on the user: Cheverst et al. [38] proposed *GUIDE*, a context-aware tourist guide. Their system relied on both personal (such as the user's interests) and environmental (such as opening time of the attraction) contexts. *GUIDE*'s users showed a high acceptance of the adaptive system, but the authors realised that it should allow a choice of functionality level as some users found *GUIDE* somehow confusing. Likewise, Kortuem et al. [117] proposed an adaptive wearable system that adapted itself based on the local environment, by communicating with intelligent objects nearby via infrared beams.

With simple approach of adaptation, McGrenere et al. [129] tested the acceptance and effectiveness of the personalisation of a complex software: users were able to chose the elements from the Full Interface they wanted to keep, and therefore two interfaces were used (Personal and Full). They reported that users appreciated the possibility to personalise the software interface, but that the system could offer a smarter way of setting the personalisation by assisting the user to create its profile ("*mixed-initiative interface*").

Gajos et al. [73] evaluated different adaptive graphical interfaces based on existing work (*Split Interface*, *Moving Interface* and *Visual Popout Interface*). Their evaluation showed that the acceptance and the performance of an adaptive interface by the users seems to

`StretchText` (last accessed Jan. 2020)

depend on the accessibility: more accesses brought dissatisfaction and low performance. They also indicated that if the frequency of adaptation is too high during a session, the users may feel unsatisfied.

Adaptation can also be retrieved from sensors that reveal the user's cognitive state or emotions. Tan and Nijholt [184] investigated the role of Brain-Computer Interaction in Human-Computer Interaction, and among other applications, they encompassed it to be used in adaptive systems, for example to evaluate when to interrupt the user. Iqbal and Bailey [99] used eye tracking to estimate the tasks performed by the users and adapted the system disruption levels accordingly. Still employing eye tracking, Iqbal et al. studied the mental workload of different tasks through pupil response [100] and proposed to use their finding in attention manager applications. Duric et al. [58] suggested other examples of biological indicators of the user's cognitive state: "*facial expressions, upper-body posture, arm movements, and keystroke force*" that can be used to build an intelligent adaptive system. Rothrock et al. complemented these examples with "*a wide range of possible inputs about the user's physiological state (e.g. EEG, heart rate variability)*", and also mentioned other user's traits that can be used for adaptive systems that yet need to be assessed before the interaction (i.e. user knowledge, user personality, cogntive style).

In this thesis, we endeavour to find out if the correlation between gaze and hand can be used to understand the user's cognitive state (hesitation in particular in our case) and if so, propose to use this input to built an adaptive interface system that can respond to the user's hesitation.

3

Study Setup

This chapter presents the methodology for setting up our studies. The first part deals with eye tracking and relates to both studies covered by Chapters 4, 5 and 6. The second part focuses on the on-screen hand gestures recognition that only concerns the work of Chapter 4 (first data collection).

3.1 Eye Tracker

The procedures found in eye tracking studies for installing and preparing the devices are often the same, positioning the eye tracker, calibrating it and using it consist in the basic steps any eye tracking study must start with. This section describes the steps and methods we followed in our work for the data collection we achieved with two eye trackers: Tobii X2-60 for the first data collection (relating to Chapter 4) and Tobii EyeX for the second data collection (relating to Chapters 5 and 6). Whenever the case happens, we will indicate in this section if we had not followed the standard procedure and why. The type of eye tracker we mention in this section are remote infrared eye trackers, such as the two eye trackers we used in our work.

3.1.1 Eye Tracker Installation

The standard configuration for remote eye tracking assumes the subject to face a vertical (or near to vertical) display (such as a computer monitor) under which the eye tracker

is attached (horizontally centred), therefore working with an image of the eyes captured from a bottom-top angle (cf. Figure 3.1). The subject can either stand up or seat, depending on the purpose of the study and the height of the monitor.

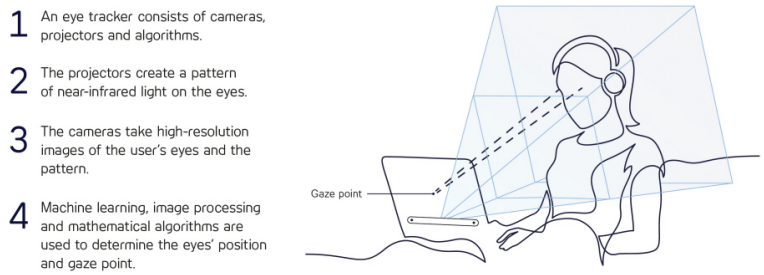


Figure 3.1: Usual remote eye tracker configuration.

Source: Tobii online documentation

However, in our work, we did not follow the usual installation of the eye tracker: our monitors (tablets) were not standing vertically but near to horizontally. In the first data collection, the subject sat on a chair and the eye tracker was placed under the tablet level (cf. Chapter 4) because (1) we used a rack especially designed for smaller displays which presents the tablet to the users in a near to horizontal position and (2), tablets are often handled near to horizontally by users when they need to interact with them manually. For the second data collection (cf. Chapters 5 and 6), the tablet was also near to horizontal, but the eye tracker was exceptionally placed **above** the screen to compensate the posture of the subjects: they stood up during interaction and a low position of the eye tracker resulted in poor eye tracking.

For both eye trackers we used, an extra step was necessary after installing the eye tracker: configuring the eye tracker with the monitor's position in space. This step is, however, not always required when using an eye tracker. Commercial eye trackers, such as Tobii EyeX are sold to be used by the general public and work with a good enough estimation of the screen dimensions based on the default spatial configuration of the device with the monitor and the calibration[1] process. In our case, we needed to inform the eye tracker of the monitor's position because of an unusual configuration. For Tobii X2-60 (used in the first data collection), Tobii provides a tool illustrated by Figure 3.2 that communicates with the eye tracker so the dimensions and position of the monitor on the rack can be sent to the eye tracker. The values that are not related to the screen dimensions were provided by Tobii with the rack's manual.

[1]The calibration is described in the next section.

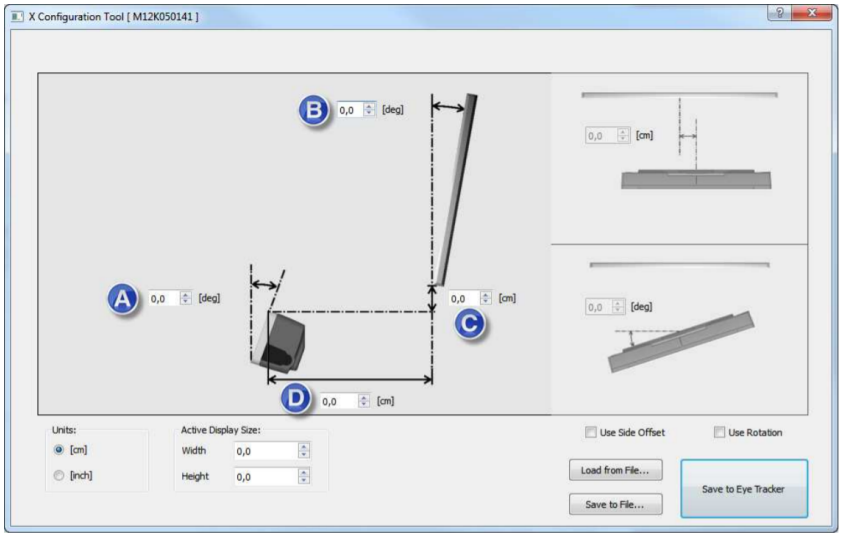


Figure 3.2: Tobii X2-60 configuration tool.

With Tobii EyeX, we used the Tobii API from a custom programme (developed in C#) allowing to provide the eye tracker with the relative position of the monitor in space. The programme snippet, illustrated by Figure 3.3, shows how this configuration is achieved: the position of three of the display corners (top left, top right and bottom left) are sent to the eye tracker via the API.

3.1.2 Eye Tracker Calibration

In order to model the eye adequately for a specific user, the eye tracker need to be calibrated. The procedure for the calibration is common among eye tracking studies: several targets are displayed on the screen sequentially, and locations which cover the whole screen area, the subject is asked to stare at them so that the eye tracker acquires data for this specific point and later, once all the targets have been displayed, compute the model of the eyes at the different points of gaze associated with the targets. When designing an application requiring an eye tracker, there is therefore a trade-off to make between usability and accuracy. A calibration with many points increases the accuracy of the model made by the eye tracker, but costs time for the user and downgrades the user experience. Besides, over calibrating is not necessarily guarantying a better interaction since it does not bring better data quality. In a research studies, 5 or 9 points are usually found for the calibration[2]. We used 9 point in our first data collection, and 5 in our second data collection.

[2]Tobii Pro SDK documentation indicates that "*Usually 5 points yields a very good result and is not experienced as too intrusive by the user.*" `http://developer.tobiipro.com/commonconcepts/calibration.html` (last accessed Jan. 2020)

```
private void configureETButton_Click(object sender, EventArgs e)
{
    //the angle INSERT in the software is the angle between the eye tracker front surface and the alignement with the tablet
    double etAngleAbs = ((double)eyeTrackerAngleAbsNumericUpdown.Value * Math.PI) / 180;

    //top left screen point
    double tlX = -SCREEN_WIDTH_MM / 2;
    double tlY = TABLET_HEIGHT_BEZEL_MM * -Math.Cos(etAngleAbs) - ET_HEIGHT / 2;
    double tlZ = TABLET_HEIGHT_BEZEL_MM * Math.Sin(etAngleAbs);
    //top right
    double trX = SCREEN_WIDTH_MM / 2;
    double trY = TABLET_HEIGHT_BEZEL_MM * -Math.Cos(etAngleAbs) - ET_HEIGHT / 2;
    double trZ = TABLET_HEIGHT_BEZEL_MM * Math.Sin(etAngleAbs);
    //bottom left
    double blX = -SCREEN_WIDTH_MM / 2;
    double blY = (TABLET_HEIGHT_BEZEL_MM + SCREEN_HEIGHT_MM) * -Math.Cos(etAngleAbs) - ET_HEIGHT / 2;
    double blZ = (TABLET_HEIGHT_BEZEL_MM + SCREEN_HEIGHT_MM) * Math.Sin(etAngleAbs);

    //set the eye tracker accordingly
    try
    {
        EyeTracker eyeTracker = new EyeTracker(new EyeTrackerCoreLibrary().GetConnectedEyeTracker());
        eyeTracker.RunEventLoopOnInternalThread(OnGenericOperationCompleted);
        eyeTracker.Connect();
        eyeTracker.SetDisplayArea(new DisplayArea(
            new Point3D(tlX, tlY, tlZ),
            new Point3D(trX, trY, trZ),
            new Point3D(blX, blY, blZ)));
        eyeTracker.Disconnect();
        eyeTracker.Dispose();
        MessageBox.Show("Display Area has been set up.");
    }
    catch (Exception ex)
    {
        MessageBox.Show(string.Format("Problem while setting the display area ({0})", ex.Message));
    }
}
```

Figure 3.3: Tobii EyeX configuration via API.

The calibration process, as well as the checkup of the calibration validity, are often simply achieved by the tools proposed the eye tracker's constructor - saving the designer of an application using eye tracking to implement these. Nevertheless, other methods may be employed instead: for the calibration process, the eye tracker's API can be used in a custom programme (which we did for our work); for the calibration validity, visual assessment can be performed by a simple control task (showing the point of gaze point on the display and asking the subject to stare at specific targets - option we chose in our first data collection) or indicators can be computed based on the eye tracker estimations and target positions (choice made in the second data collection). These indicators are the *precision* and the *accuracy*. The notions of precision and accuracy are illustrated in Figure 3.4: a precise calibration is consistent in the estimation samples' value (for a given target the estimations are not varying a lot), an accurate calibration shows estimation samples reliably close to the ground truth.

Precision and accuracy are, for eye tracking, expressed in degrees of visual angle, and respectively computed as showed by equations 3.1 (where n is the number of samples, i is the 0-based index of a sample and θ is the value of visual angle between sample i and sample *i-1*) and 3.2 (where n is the number of samples, i is the 0-based index of a sample and α is the value of visual angle between sample i and the ground truth target). Acceptable values for accuracy and precision depend on the purpose the eye tracker. For fixations and saccades detection, a precision value up to 0.05° is commonly accepted,

while accuracy is less critical and can reach up to 1° [94].

$$precision = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}\theta_i^2} \tag{3.1}$$

$$accuracy = \frac{1}{n}\sum_{i=0}^{n-1}\alpha_i \tag{3.2}$$

Over time, eye trackers often need to be recalibrated. Factor for such drift are explained by Holmqvist et al. in [94] as a result of "*physical conditions change after calibration*". It is therefore custom, in research work with eye tracking, to report the *drift* observed after a data collection or experiment session. The procedure to measure drift is the same as the calibration.

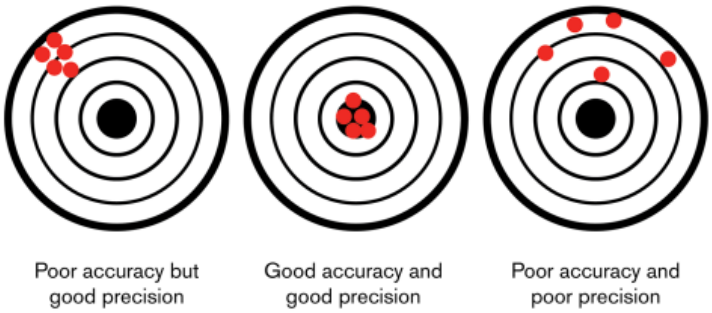


Figure 3.4: Eye tracker precision and accuracy.

Source: Tobii Pro online documentation

3.1.3 Eye Tracker Data Interpretation

The samples provided by the eye trackers we used contain a time stamp, a validity code and the estimated gaze of point for each eye. The eye tracker API usually contains methods to synchronise the eye tracker and the computer clocks, as well as to retrieve time stamps that take the latency of the system into account. It is possible that at a moment in time during interaction, the eye tracker only tracks one eye, both eyes or none during interaction (for instance, when blinking). Such loss of data is expected and when eye trackers are used for research purpose, data from participants which presents too much loss should be discarded. There is no specification related to the thresholds or criteria

that should be observed in order to consider data from a participant valid. Literature usually provides this information on a study basis with the authors' appreciation. For our work, we based the threshold of data quality upon two different criteria. We used a comparison criterion for the first study between the number of touch input samples and the number of gaze samples during the completion of a task. The eye tracker frequency was 60 Hz and touch events was approximately 100 Hz. So, the ideal ratio touch/eye is about 1,66666. If the ratio for a task was more than 2 then it was considered the eye data was not sufficient. For the second study, the criterion is the duration of the task. Knowing the eye tracker's frequency, it is easy to check what percentage of samples are valid for a given period. Our threshold values was 60 %.

3.1.4 Eye Tracker Limitations

Because eye trackers work with a natural input (the eyes) and suffer from technical issues, they limit the scope of experiments (distance between the user and the tracker, model of the eye, lightening environment ...).

Remote infrared eye trackers are commonly used in eye tracking, such as Tobii X-20 and Tobii EyeX. They emit infrared on the eye, and a combined image of the eye balls in natural and in infrared lights is then analysed against an eye model to estimate the point of gaze. Therefore, these eye trackers have several inherent constraints. First, the external environment where the eye tracker is used must avoid sources of infrared lights that could create an interference with the eye tracker system (for example, other sensor beams, sunlight...). Makeup can also interfere with eye tracking (because of reflective substances) and it is usually advised to ask eye tracking subjects not to wear makeup. The subject must also stand in front of the eye tracker in a distance and a position that allow the eyes to be tracked. Moreover, eye tracking does not permit several users to interact simultaneously with an eye tracker: only one subject's eyes can be tracked by one eye tracker. Personal features may also impact the quality of eye tracking as the eyes of the subject may not be modelled properly under some conditions. The model of the eye relies on standard eye characteristics that implies the subjects have healthy eyes (shape that stays in the limits allowed by the model, detectable pupil etc). Most eye trackers (at least the ones we used in our work) can model the eye even if the subject wears glasses or contact lenses - as long as the lenses do not distort the image of the eye and the infrared light strongly. Therefore, some type of eye correction may prevent from the eye modelling to be perform well and the eye tracker to be used. The eye lid also

plays a role in the eye detection: droopy eye lid may cover the eyes too much for the eye tracker to be able to model the eye correctly.

3.2 On-Screen Hand Gesture Detection

On-screen hand gesture recognition was required for the first data collection we made (covered by Chapter 4). In this section, we present the technical choices and principal steps that we implemented to collect manual hand input from a tablet and get the gesture associated with this input. The design of our application restricted the expected gestures to taps, pans (drag) and, potentially, zooms. In the following, we present the two complementary solutions we employed for (1) collecting raw input data and (2) interpret the data to retrieve the hand gestures performed on the tablet.

3.2.1 Raw Manual Input Data

In order to process the data from the on-screen manual input, we looked at how to retrieve this input directly from the digitizer itself. To achieve so, we initially parsed the HID reports of the tablet's embedded screen, using the native Windows Raw Input API. Human Interface Device (HID) reports are standardised files that a device presents to the operating system to describe how its data should be interpreted by the latter.

Microsoft provides online documentation regarding how to search for the appropriate data (of the digitizer) in the HID reports: the data can be reached by finding the right *page* and *usage* in the report. For instance, the X and Y values of a contact point in the digitizer is found at the page *Desktop* (0x01) at the usages 0x30 and 0x31. Documentation on the functions and types exposed by other native APIs we used is also available online.

The following describes the main steps to retrieve the raw input data from the digitizer using the Raw Input API of Microsoft. We wrote a C# programme to call the native API functions and parse the data. The API, as well other native Microsoft APIs are linked with the C# programme via the following dll files: `user32.dll` for the Raw Input API, `hid.dll` for the system API to work with HID reports, `setupAPI.dll` for getting information about the devices managed by the OS and `kernel32.dll` for access to the OS kernel, memory and system resources.

Firstly, the digitizer (as registered in the HID list) need to be found. Figure 3.5 shows the code snippet to retrieve the digitizer (API function `GetRawInputDeviceInfo`) and

register it with the API to lookup its report (API function `RegisterRawInputDevices`). The *vendor* and *product* identifiers are found against the HID report looked up at the page *Digitizer* (0x0D) for the usage *Joystick* (0x04).

```
public static bool GetRawInputDeviceList()
{
    bool res = false;
    RAWINPUTDEVICELIST_ELMT[] pRawInputDeviceList = null;
    uint puiNumDevices = 0;
    uint returnCode = GetRawInputDeviceList(pRawInputDeviceList, ref puiNumDevices, (uint)Marshal.SizeOf(new RAWINPUTDEVICELIST_ELMT()));
    res = (0xFFFFFFFF != returnCode);
    if (res)
    {
        //alloc array
        pRawInputDeviceList = new RAWINPUTDEVICELIST_ELMT[puiNumDevices];

        //get devices
        returnCode = GetRawInputDeviceList(pRawInputDeviceList, ref puiNumDevices, (uint)Marshal.SizeOf((typeof(RAWINPUTDEVICELIST_ELMT))));
        res = (0xFFFFFFFF != returnCode);
        if (res)
        {
            //look for the touchscreen.
            bool foundTouchScreen = false;
            foreach (RAWINPUTDEVICELIST_ELMT rawInputDevice in pRawInputDeviceList)
            {
                uint structsize = (uint)Marshal.SizeOf(typeof(RID_DEVICE_INFO));
                RID_DEVICE_INFO di = new RID_DEVICE_INFO();
                di.cbSize = structsize;
                IntPtr pData = Marshal.AllocHGlobal((int)structsize);
                returnCode = GetRawInputDeviceInfo(rawInputDevice.hDevice, RawInputDeviceInfoType.RIDI_DEVICEINFO, pData, ref structsize);
                if (0xFFFFFFF != returnCode && 0 != returnCode)
                {
                    di = (RID_DEVICE_INFO)Marshal.PtrToStructure(pData, typeof(RID_DEVICE_INFO));
                    switch (di.dwType)
                    {
                        case RawInputDeviceType.RIM_TYPEHID:
                            if (HID_USAGE_PAGE_DIGITIZER == di.hid.usUsagePage && HID_USAGE_GENERIC_JOYSTICK == di.hid.usUsage)
                            {
                                _vendorID = di.hid.dwVendorId;
                                _productID = di.hid.dwProductId;
                                foundTouchScreen = true;
                            }
                            break;
                        case RawInputDeviceType.RIM_TYPEKEYBOARD:
                        case RawInputDeviceType.RIM_TYPEMOUSE:
                        default:
                            break;
                    }
                    if (foundTouchScreen)
                    {
                        RAWINPUTDEVICE[] rawInputDevicesToMonitor = new RAWINPUTDEVICE[1];
                        rawInputDevicesToMonitor[0].UsagePage = di.hid.usUsagePage;
                        rawInputDevicesToMonitor[0].Usage = 0;
                        rawInputDevicesToMonitor[0].Flags = (int)(RawInputDeviceFlags.InputSink | RawInputDeviceFlags.PageOnly);
                        rawInputDevicesToMonitor[0].WindowHandle = _associatedForm.Handle;

                        if (!RegisterRawInputDevices(rawInputDevicesToMonitor, (uint)rawInputDevicesToMonitor.Length, (uint)Marshal.SizeOf(rawInputDevicesToMonitor[0])))
                        {

                            _associatedForm.WriteLogInForm("Registration of device --> NOK (error: " + Marshal.GetLastWin32Error() + ")");
                        }
                        else
                        {
                            FindHidDevice(_vendorID, _productID, out _touchScreenDevice);
                        }
                        break;
                    }
                }
            }
        }
    }
    return res;
}
```

Figure 3.5: Code snippet showing how to retrieve and register the digitizer with the Raw Input API.

Our C# application then overrides the function `WndProc`. Microsoft defines it as "*An application-defined function that processes messages sent to a window.*" (online documentation). When this callback method is called, we set up a time stamp of the touch event, and we check whether the message sent as an argument to the method is of the right type (API type `WM_INPUT`) and attempt to read the HID report data. The data is obtained by reading the values of the desired pages/usages. An example of reading values is shown in the code snippet illustrated by Figure 3.6: the API function `HidP_GetUsageValue` is called whenever a value need to be retrieved.

Whenever the C# application receives an HID report, it parses it, then prepares and sends a data packet into a network socket over TCP, on the same machine (local address `127.0.0.1`, port 5945) where the application runs (the tablet). If several contact points

```
public static RawInputDataEx ReadHIDData(ref System.Windows.Forms.Message m, DateTime dtEvent)
{
    RawInputDataEx rawInputDataEx = new RawInputDataEx();
    rawInputDataEx.Timestamp = dtEvent;

    uint returnCode;
    IntPtr pRawInputData = IntPtr.Zero;
    uint pcbSize = 0;
    returnCode = GetRawInputData(m.LParam, RawInputDataCmd.RID_INPUT, pRawInputData, ref pcbSize, (uint)Marshal.SizeOf(new RAWINPUTHEADER()));
    if (0xFFFFFFFF != returnCode)
    {
        //data is received
        int sizeOfRawInputHeader = Marshal.SizeOf(new RAWINPUTHEADER());
        pRawInputData = Marshal.AllocHGlobal((int)pcbSize);
        returnCode = GetRawInputData(m.LParam, RawInputDataCmd.RID_INPUT, pRawInputData, ref pcbSize, (uint)sizeOfRawInputHeader);
        if (0xFFFFFFFF != returnCode)
        {
            RAWINPUT rawInputData = (RAWINPUT)Marshal.PtrToStructure(pRawInputData, typeof(RAWINPUT));
            if (rawInputData.header.dwType == RawInputDeviceType.RIM_TYPEHID)
            {
                if (rawInputData.hid.dwSizeHid > 0)
                {
                    //because of the limitation in the struct definition (no array) we look for the byte[] directly in memory.
                    byte[] rawHidData = new byte[rawInputData.hid.dwCount * rawInputData.hid.dwSizeHid];
                    for (int i = 0; i < rawHidData.Length; i++)
                    {
                        //reach the data : we know the pRawInputData is a RAWINPUT +  data; made of header (4*4 bytes) and RAWHIDNoData part (2*4 bytes)
                        rawHidData[i] = Marshal.ReadByte(IntPtr.Add(pRawInputData, sizeOfRawInputHeader + 8 + i));
                    }

                    //use the Hid routines to find the data according to the device's parsed reports.
                    bool isDeviceRetrieved = (_touchScreenDevice.pHidDevice != IntPtr.Zero);
                    if (!isDeviceRetrieved)
                        isDeviceRetrieved = FindHidDevice(_vendorID, _productID, out _touchScreenDevice);
                    if (isDeviceRetrieved)
                    {
                        IntPtr reportPtr = Marshal.AllocHGlobal(_touchScreenDevice.Caps.InputReportByteLength);
                        Marshal.Copy(rawHidData, 0, reportPtr, rawHidData.Length);
                        IntPtr usageValuePtr = Marshal.AllocHGlobal(4);

                        //Contact counts (page 0xD, usage 0x54, (taille 1x1bytes - reportID 3, reportSize 27, in VALUE caps)) (we also know the max counts is 10 on our device)
                        returnCode = HidP_GetUsageValue(HIDP_REPORT_TYPE.HidP_Input, HID_USAGE_PAGE_DIGITIZER,
                            0, 0x54, usageValuePtr, _touchScreenDevice.Ppd, reportPtr, _touchScreenDevice.Caps.InputReportByteLength);
                        if (HIDP_STATUS_SUCCESS == returnCode)
                        {
                            byte[] data = new byte[1];
                            Marshal.Copy(usageValuePtr, data, 0, 1);
                            rawInputDataEx.ContactsCount = (ushort)data[0];

                        }
                        else
                        {
                            returnCode = 0xFFFFFFFF;
                            goto Cleaning;
                        }

                        //X (page 0x1, usage 0x30, (taille 2x2bytes - reportID 3, reportSize 27, in VALUE caps))
                        returnCode = HidP_GetUsageValueArray(HIDP_REPORT_TYPE.HidP_Input, HID_USAGE_PAGE_GENERIC,
                            0, HID_USAGE_GENERIC_X, usageValuePtr, 4, _touchScreenDevice.Ppd, reportPtr, _touchScreenDevice.Caps.InputReportByteLength);
                        if (HIDP_STATUS_SUCCESS == returnCode)
                        {
                            //return order is little endian
                            byte[] data = new byte[4];
                            Marshal.Copy(usageValuePtr, data, 0, 4);
                            rawInputDataEx.Xt = (ushort)(data[0] + 256 * data[1]);
                            rawInputDataEx.Xc = (ushort)(data[2] + 256 * data[3]);
                            if (_touchScreenWidth <= 0)
                            {
                                foreach (HIDP_VALUE_CAPS valCap in _touchScreenDevice.pInputValueCaps)
                                {
                                    if (valCap.UsagePage == HID_USAGE_PAGE_GENERIC)
                                    {
                                        ushort valUsage;
                                        if (valCap.IsRange)
                                        {
                                            valUsage = valCap.Range.UsageMax;
                                        }
                                        else
                                        {
                                            valUsage = valCap.NotRange.Usage;
                                        }
                                        if (valUsage == HID_USAGE_GENERIC_X)
                                        {
                                            _touchScreenWidth = valCap.LogicalMax;
                                            break;
                                        }
                                    }
                                }
                            }
                            rawInputDataEx.Width = _touchScreenWidth;
                        }
                        else
                        {
                            returnCode = 0xFFFFFFFF;
                            goto Cleaning;
                        }

                        //Y (page 0x1, usage 0x31, (taille 2x2bytes - reportID 3, reportSize 27, in VALUE caps))
                        returnCode = HidP_GetUsageValueArray(HIDP_REPORT_TYPE.HidP_Input, HID_USAGE_PAGE_GENERIC,
                            0, HID_USAGE_GENERIC_Y, usageValuePtr, 4, _touchScreenDevice.Ppd, reportPtr, _touchScreenDevice.Caps.InputReportByteLength);
                        if (HIDP_STATUS_SUCCESS == returnCode)
                        {
                            //return order is little endian
                            byte[] data = new byte[4];
                            Marshal.Copy(usageValuePtr, data, 0, 4);
                            rawInputDataEx.Yt = (ushort)(data[0] + 256 * data[1]);
                            rawInputDataEx.Yc = (ushort)(data[2] + 256 * data[3]);
                            if (_touchScreenHeight <= 0)
                            {
                                foreach (HIDP_VALUE_CAPS valCap in _touchScreenDevice.pInputValueCaps)
                                {
                                    if (valCap.UsagePage == HID_USAGE_PAGE_GENERIC)
                                    {
                                        ushort valUsage;
                                        if (valCap.IsRange)
                                        {
                                            valUsage = valCap.Range.UsageMax;
                                        }
                                        else
                                        {
                                            valUsage = valCap.NotRange.Usage;
                                        }
                                        if (valUsage == HID_USAGE_GENERIC_Y)
                                        {
                                            _touchScreenHeight = valCap.LogicalMax;
                                            break;
                                        }
                                    }
                                }
                            }
                            rawInputDataEx.Height = _touchScreenHeight;
                        }
                        else
                        {
                            returnCode = 0xFFFFFFFF;
                            goto Cleaning;
                        }

                        //Contact ID (page 0xD, usage 0x51, (taille 1x2bytes - reportID 3, reportSize 27, in VALUE caps))
                        returnCode = HidP_GetUsageValue(HIDP_REPORT_TYPE.HidP_Input, HID_USAGE_PAGE_DIGITIZER,
                            0, 0x51, usageValuePtr, _touchScreenDevice.Ppd, reportPtr, _touchScreenDevice.Caps.InputReportByteLength);
                        if (HIDP_STATUS_SUCCESS == returnCode)...
                        else...

                    //cleaning
                    Cleaning:
                        Marshal.FreeHGlobal(reportPtr);
                        Marshal.FreeHGlobal(usageValuePtr);
                    }
                    else
                    {
                        returnCode = 0xFFFFFFFF;
                    }

                }
                else
                {
                    returnCode = 0;
                }
            }
        }
        Marshal.FreeHGlobal(pRawInputData);
    }
    rawInputDataEx.GetDataReturnCode = returnCode;
    return rawInputDataEx;
}
```

Figure 3.6: Code snippet showing how to read the HID reports data of the digitizer with the Raw Input API.

are sensed by the digitizer, the API returns a message serially per contact point. We gather them in one single packet that contains the time stamp (8 bytes, a tick value as implemented by C# in the `DateTime` class), the number of contact points (1 byte) and for each contact points: the point ID (4 bytes), the normalised X coordinate (4 bytes), the normalised Y coordinate (4 bytes) and the touch point status (1 byte, value of 0 if the contact point is at the birth status, 1 if the contact point is at the death status, and 2 if the contact point is at the move status). All information is encoded in big-endian order.

3.2.2 Third Party Gesture Detection

The Raw Input API described in the previous section does not inform of the on-screen gestures directly. Therefore another tool is necessary to analyse the raw input data and return the right gesture. We used a third party application, Sparsh UI[3] that can do so when raw input data is sent over via a network socket. Sparsh UI is a Java application that consists of a server to receive the raw data, prepare it for interpretation and send the data to a client that interprets this data as on-screen gesture. In our case, the C# application send the raw data as explained in Section 3.2.1 to the Sparsh UI server, and we used the client API of Sparsh UI to write file logs.

```
private void readTouchPoint(long timeStamp) throws IOException {
    int id = _in.readInt();
    float x = _in.readFloat();
    float y = _in.readFloat();
    Location location = new Location(x, y);
    TouchState state = TouchState.values()[(int)_in.readByte()];
    processTouchPoint(id, location, state, timeStamp);

    // DEBUG
    System.out.println("[InputDeviceConnection] ID: " + id + " STATE: " +state.name() +
            " X: " + x + " Y: " + y + " TIMESTAMP: " + timeStamp);

    if (state == TouchState.DEATH) {
        flagTouchPointForRemoval(id);
    }
}

private void readTouchPoints() throws IOException {
    //PWT read the time stamp
    long timestampTick = _in.readLong();

    //PWT changed the reading method to read a value on a byte instead of 4 bytes originally
    int count = _in.readUnsignedByte();
    if(count < 0) {
        _in.close();
        return;
    }
    for(int i = 0; i < count; i++) {
        readTouchPoint(timestampTick);
    }
    removeDeadTouchPoints();
}
```

Figure 3.7: Code snippet showing how the raw input is processed by Sparsh UI.

[3] `https://code.google.com/archive/p/sparsh-ui/` (last accessed Jan. 2020)

```
public TouchEventGestureClient() {
    // prepare the allowed gesture
    _allowedGesture = new Vector<Integer>();
    _allowedGesture.add(new Integer(GestureType.DRAG_GESTURE.ordinal()));
    //_allowedGesture.add(new Integer(GestureType.MULTI_POINT_DRAG_GESTURE.ordinal()));
    //_allowedGesture.add(new Integer(GestureType.ROTATE_GESTURE.ordinal()));
    _allowedGesture.add(new Integer(GestureType.TOUCH_GESTURE.ordinal()));
    _allowedGesture.add(new Integer(GestureType.ZOOM_GESTURE.ordinal()));
    //_allowedGesture.add(new Integer(GestureType.DBLCLK_GESTURE.ordinal()));
    //_allowedGesture.add(new Integer(GestureType.FLICK_GESTURE.ordinal()));
    //_allowedGesture.add(new Integer(GestureType.RELATIVE_DRAG_GESTURE.ordinal()));

    // connect to the server
    try {
        new ServerConnection("127.0.0.1", this);
    } catch (UnknownHostException e) {
        System.out.println("[TouchEventGestureClient] Server Host Unknown");
        e.printStackTrace();
    } catch (IOException e) {
        System.out
                .println("[TouchEventGestureClient] Failed to establish server connection");
        e.printStackTrace();
    }
}
```

Figure 3.8: Code snippet showing what selected gestures are interpreted by Sparsh UI.

At establishing the first connection with the Sparsh UI server, the C# application of Section 3.2.1 sends a single byte (value of 1) as required by Sparsh UI to register the new sender. We consecutively also send extra information related to our study work: the participant index and the task type, as these elements are saved in the log file's name. The server then enters into a loop mode to receive the data packets from the network socket. Figure 3.7 shows the part of the server that extract the time stamp (added for the purpose of our work) and the raw input data (via the method `readTouchPoints` of the class `InputDeviceConnection`).

Sparsh UI's gesture detection logic starts with the method `processTouchPoint` of the class `InputDeviceConnection`. We altered the method by adding an extra argument for the time stamp, which is sent through all other subsequent methods. Also, we limited the gestures recognition to taps, drags and zoom events in the constructor of the class `TouchEventGestureClient` as illustrated by Figure 3.8.

Sparsh UI handles the gesture recognition by performing several check points[4] for each type of gestures allowed by Sparsh UI against the touch data of a same group. A group, as seen by Sparsh UI, gathers several touch points and is alive as long as at least one identified touch of the group is still alive: for example, if during interaction the user touched the tablet with a finger (assuming there was no interaction), Sparsh UI would keep the group alive as long as at least a touch exists without any break at any moment in time.

[4]We do not, in this thesis, intend to explain the details of Sparsh UI internal logic. As an example: for the zoom gesture, a check point can be the presence of two simultaneous touch points.

```
@Override
public void processEvent(int groupID, Event event) {
    PrepareLogFile();
    // we process the event as a logging method --> event is not sent to any
    // application
    EventType eventType = EventType.values()[event.getEventType()];
    float coordX=-1;
    float coordY=-1;
    String param1="-";

    boolean writeLog=true;

    switch (eventType) {
    case DRAG_EVENT:
        DragEvent dragEvent=(DragEvent) event;
        coordX=dragEvent.getAbsX();
        coordY=dragEvent.getAbsY();
        writeLog = (coordX !=0 || coordY!=0);
        break;
    case TOUCH_EVENT:
        TouchEvent touchEvent=(TouchEvent) event;
        coordX=touchEvent.getX();
        coordY=touchEvent.getY();
        break;
    case ZOOM_EVENT:
        ZoomEvent zoomEvent=(ZoomEvent) event;
        coordX=zoomEvent.getCenter().getX();
        coordY=zoomEvent.getCenter().getY();
        param1=Float.toString(zoomEvent.getScale());
        writeLog=(zoomEvent.getScale()!=1);
        break;
    default:
        System.out.println("[TouchEventGestureClient] Unknow event type.");
        try {
            _fileWriter.write("[TouchEventGestureClient] Unknow event type.\n");
        } catch (IOException e) {
            System.out.println("[TouchEventGestureClient] Failed to write log in the log file.");
        }
    }

    if(writeLog){
        // open the log file
        try {
            _fileWriter = new FileWriter(_touchLogFile, true);
        } catch (IOException e1) {
            System.out
                    .println("[TouchEventGestureClient] Log file can not be used");
            e1.printStackTrace();
        }

        // CSV FILE HAS THIS FORMAT :
        // timeStamps;event;X;Y;param1
        try {
            _fileWriter.append(String.format("%d\t%s\t%f\t%f\t%s\n",
                    event.getEventTimeStamp(), eventType, coordX, coordY,
                    param1));
        } catch (IOException e) {
            System.out
                    .println("[TouchEventGestureClient] Failed to write log in the log file.");
            e.printStackTrace();
        }

        try {
            _fileWriter.close();
        } catch (IOException e) {
            System.out
                    .println("[TouchEventGestureClient] Failed to close log file.");
            e.printStackTrace();
        }
    }
}

private void PrepareLogFile() {
    _touchLogFile = "C:\\TabletHandEyeCorrelationStudy\\DATA\\Touch_" + _participantID + "_" + _taskType +".tsv";
    File file=new File(_touchLogFile);
    if(!file.exists()){
        try {
            FileWriter fw = new FileWriter(file);
            fw.write("timeStampClicks\tEvent\tX\tY\tparam1\n");
            fw.close();
        } catch (IOException e) {
            System.out.println("[TouchEventGestureClient] Failed to open, write or close log in the log file.");
        }
    }
}
```

Figure 3.9: Code snippet showing the writing of on-screen gestures log files with Sparsh UI.

Once the gesture has been identified, we use the class `TouchEventGestureClient` that needs to implement the API's class `Client` to write the resulting data (timestamp, gesture type, coordinates and additional parameter) into a log file. The code snippet in Figure 3.9 shows the content of the method `processEvent` in which this is achieved.

4

Correlation between Gaze and Tap

4.1 Introduction

When interacting with computing devices, manual input is highly connected with how users visually inspect UI content. The common Human-Computer Interaction scenario implicates the hand as an (indirect) input and the eyes as the monitoring element (to analyse the interface). The correlation between manual input (using the mouse acting as a proxy for the hand) and gaze has been of particular interest in many research efforts [15, 33, 123], to better understand visual attention across the variety of computing devices we use on a daily basis, and to propose new concepts that enhance the interaction. However, besides the increasing popularity of tactile devices, correlation between touch input and gaze has, to our knowledge, not been studied yet.

In this work, we investigate how tapping correlates with gaze on a tablet device. We conducted a study with 24 participants and collected data related to touch input, gaze and tapped targets. Our data collection context turns to Internet related tasks, because browsing is a typical task widely used as study context for measuring mouse-eye correlation [97] and commonly performed by tablet users.

Analysis of the data indicates the following results: (1) gaze preceded touch with similar spatial and temporal features than observed with the mouse, and (2) the distance kept between the gaze and the touch varies across users, and was influenced by the learning and anticipation effects of the tasks.

4.2 Pilot study

4.2.1 Introduction

In the designing process of the data collection to study gaze and tap correlation on a tablet, we considered setting up a commercial rack from Tobii, especially made for interaction with small touch devices such as phones and tablets. However, the presence of the guide bars that were part of the rack (cf. Section 4.3) raised concerns over keeping the naturalness of the participants' interaction during the data collection. In other words, we wondered whether the guide bars would prevent the participants from tapping onto the tablet as they would do so in a natural context, when no other equipment accompanies the tablet.

4.2.2 Method and Apparatus

To answer the question of the guide bars' influence on the tablet interaction, we conducted a short pilot study with 6 participants (recruited among other students of the School of Computing and Communications at Lancaster University - 1 female, 1 left handed, age 28.5±9.1) to perform a target selection task (by tapping) on a tablet (Microsoft Surface Pro 3, landscape mode, screen size 12 inches, resolution 2160×1440) under two conditions.

For the first condition, the tablet was mounted on the Tobii rack (with the guide bars), whereas for the second condition, the tablet was left stand-alone on the desk (in a position close to horizontal to avoid the tablet to slide when tapping, and also to keep a similar position of the tablet with the rack condition).

For both setups, the participants were instructed to tap on a red circle-shaped target as fast as possible with their dominant hand. In total, 15 targets sequentially appeared on the screen, in the same pseudo random order for all participants, as illustrated by Figure 4.1. The targets of 22 pixels radius were distributed across the tablet's display (3 rows, 5 columns). They were placed at the following screen ratios $[0.1, 0.3, 0.5, 0.7, 0.9] \times [0.1, 0.5, 0.9]$. To trigger the sequence, participants had to tap on the first target which was displayed right after the application was launched. For all targets, if a tap was outbound, the sequence did not continue and the target remained shown. Each participant performed 10 trials on each setup (tablet on Tobii stand or on the desk). The setup order was counterbalanced between participants.

During the trials, the following metrics were recorded: target's selection completion time, distance between tap position and the target centre, and number of outbound taps (fails).

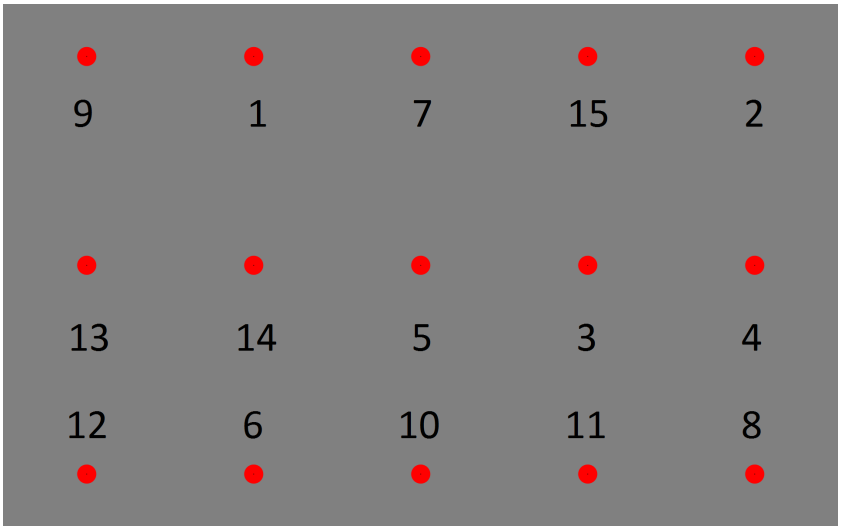


Figure 4.1: Pseudo-random sequence order of the pilot study targets.

4.2.3 Results

In the following results, *condition* refers to the experiment setup configurations we compare, where "t" is the configuration using the Tobii stand and "h" is the configuration using the tablet by itself only (home made designed configuration[1]).

4.2.3.1 Completion Time

The first target of each trial is discarded in the completion time observation, because the timer started at the very beginning of the experiment with the first target shown: participants did not have a preparatory dull tap beforehand nor signal to start on. Moreover, a dialogue box appeared between each trial. Therefore, we cannot evaluate the completion time for the first target as the tap of the first target did not comply with the same routine used for the other targets.

The completion time means are shown in Table 4.1 (means per participant and condition). We observe similar values of the mean completion time for each condition. Even if these results may imply a slightly better performance for the Tobii stand condition, the means

[1]Initially, for the "h" condition, we thought of building our own tablet support because the Tobii rack was designed for eye tracking studies, and therefore we needed something to hold the eye tracker in that condition too. However, we only wanted to investigate the impact of the guide bars on the taps, where eye tracking was unnecessary. Thus, constructing a home made rack was no longer required.

per condition and per participant (Table 4.1) demonstrate that it is not always the case for all the participants. A boxplot chart of the completion time means per target and per condition (Figure 4.2) shows how similar the completion time is for each condition.

Table 4.1: Mean completion time (per participant, per condition).

Participant	Condition	Completion Time (ms)
1	h	1101.7907
1	t	1009.9700
2	h	1351.5993
2	t	1194.8221
3	h	979.4829
3	t	1076.6386
4	h	965.9157
4	t	1013.9164
5	h	911.0043
5	t	1027.8671
6	h	891.3679
6	t	803.8886
all	h	1033.527
	t	1021.184

4.2.3.2 Distance Error

The distance error is observed within the **succeeded** taps. Based on the target's radius, the maximum distance error is 22 pixels from the target centre. The distance error is a measurement revealing how accurate is the participants' touch. Table 4.2 shows the mean distance error per participant and per condition, and among all participants. Again, obtained results indicate a similarity of the distance error between the two conditions.

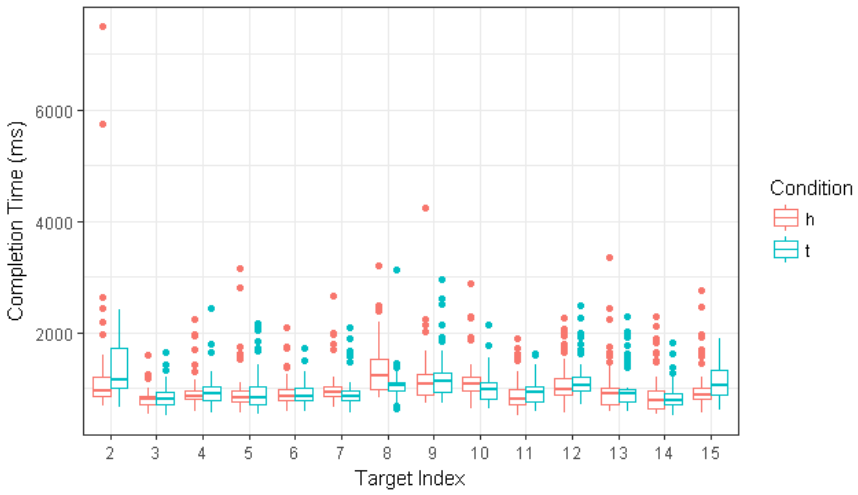


Figure 4.2: Completion time (per target, per condition).

Table 4.2: Mean distance error (per participant, per condition).

Participant	Condition	Distance Error (pixels)
1	h	10.698868
1	t	11.441963
2	h	12.473454
2	t	11.498071
3	h	12.483907
3	t	12.718841
4	h	10.408145
4	t	9.264039
5	h	12.235693
5	t	11.009178
6	h	12.369670
6	t	11.468581
all	h	11.7782
	t	11.23345

Using the Tobii rack seems to be slightly more accurate (except for two participants), but for all participants, the mean distance's difference between the two conditions is just about a pixel.

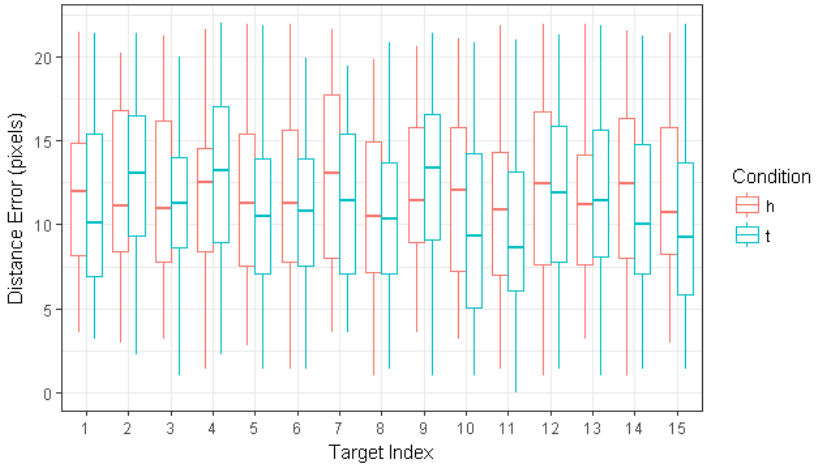


Figure 4.3: Distance error (per target, per condition).

The mean distance error boxplots per condition (Figure 4.3) does not lead to any particular conclusion.

4.2.3.3 Failed Tap Attempts

The metric of the failed tap attempts at a target acts an indicator for the difficulty to select the target. A tap attempt was failed if it occurred outside the target's 22-pixel radius bounds. We measured the number of failed tap attempts for each target. Table 4.3 illustrates the very low difference between each condition. When observing the failed

tap attempts mean per participant, we can even not conclude which condition offers better ways to complete the selections since the generation of less failures for one specific condition is equally balanced among the participants.

Table 4.3: Mean failed tap attempts (per participant, per condition).

Participant	Condition	Failed Tap Attempts
1	h	0.11
1	t	0.18
2	h	0.26
2	t	0.16
3	h	0.21
3	t	0.29
4	h	0.05
4	t	0.00
5	h	0.15
5	t	0.23
6	h	0.30
6	t	0.16
all	h	0.18
	t	0.17

The values reported in Table 4.3 also indicate that failed tap attempts are on average very low, regardless the configuration.

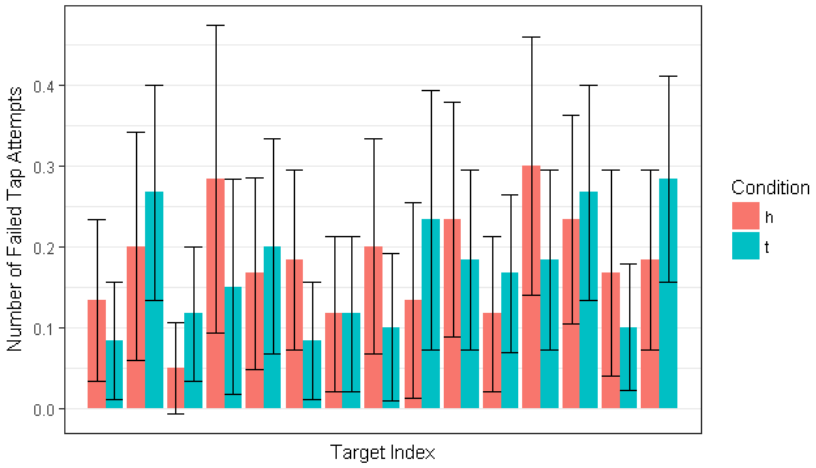


Figure 4.4: Mean failed tap attempts (per target, per condition).

4.2.4 Conclusion

This pilot study has not provided us with obvious results suggesting that we should use (or not use) the Tobii stand. Even if the means were equivalent, the Tobii condition seemed to be a slightly better condition. However, when observing the means per more

variables (per participants, per targets) for every metric observed, no obvious pattern could be found.

Therefore, we chose to conduct the main study described in the further sections using the Tobii rack. This choice was based on the lack of significant difference between the two conditions (there was no guarantee that not using the stand would bring better results), and also on the ease of integrating the gaze tracker in the apparatus environment, helping us getting more accurate and reliable gaze data.

4.3 Study Design

4.3.1 Data Collection Context

We sought to conduct a data collection based on activities that are commonly found in ordinary tasks with tablets, in order to study the correlation between gaze and tap in a natural environment. The choice of Internet activities seemed to be a good one, not only because it met this criterion, but also because research literature related to Internet based tasks in user studies is numerous. Literature therefore provided examples and inspirations on designing the data collection's contextual tasks.
We have devised three different tasks to cover the different aspects of Internet related activities on tablets.

4.3.1.1 Search Task

The search task comprised ten questions that the participants were asked to answer, by finding information on the Internet with the means of their choice. The search task is an example of a very popular activity on Internet, probably explaining why a lot of studies on Search Engine Results Pages (SERP) are found in literature. They indicate how the task can be conducted and give examples of search questions. We chose five *informal* and five *navigational* questions, since these two types of research queries are the most frequent. The questions were selected based on similar research articles, and they are listed in Appendix A, Table A.1. Google had been set as the default search engine in the browser participants worked with during the study. However, when the task started, the page appeared blank to give the participants the choice to use their preferred way of searching. The different strategies they showed during study were: either (1) reaching the Google website and typing the query in Google or (2) querying directly into the

address bar. Figure 4.5 shows a screenshot of the browser while performing the search task, resulting to an automatic query on Google by the browser.

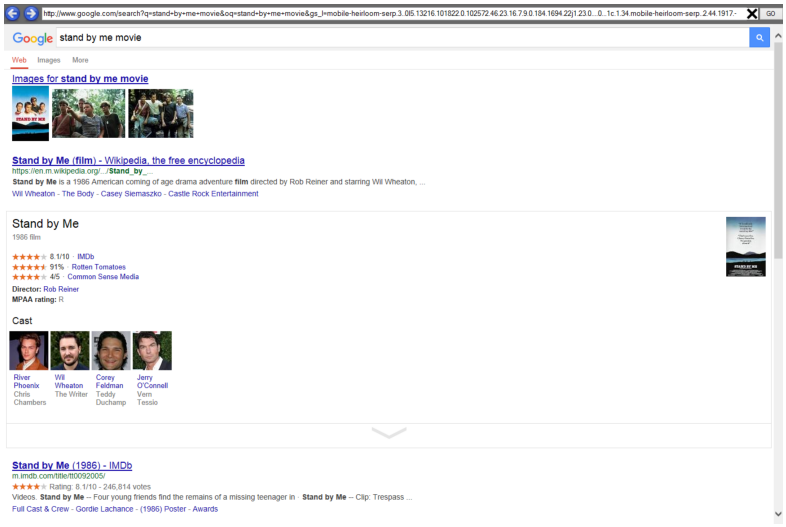


Figure 4.5: Search task.

4.3.1.2 Shopping Task

The choice of a shopping activity was not only pointing towards another very common activity on Internet, but it permitted also to bring more interaction from participants with forms (therefore including typing). Contrary to SERP activities, the websites in Shopping activities resemble an application, since the structure of the webpage is more various than the expected results list in SERP. We asked participants to simulate the purchase of items as if they were shopping to prepare a meal. We asked them to shop at least 10 different items (in order to have enough interaction with the website) and the workflow of the website we selected required the participants to fill a form. They were free to fill it with random data of their mind or to follow a guideline that contained fake personal data they could use instead (Appendix A, Table A.2). For this task, the browser started off with the Sainsbury's online groceries store website main page loaded, as illustrated by Figure 4.6.

4.3.1.3 Game Task

The game task was based on the so-called "Wikipedia game". This task has been brought on as a way to generate a productive link following activity. Moreover, this gamified task was a nice motivation to keep participants playing the game and spend time on the study. The game consisted in asking the participants to use **only** the **internal** links

of Wikipedia's articles for reaching a predefined target article from a predefined source article. Two rounds of the game could be performed by the participants (if they felt difficulty with the first round, we stopped the game or asked them to try the second round. We stressed out there was no measurement of their performance so they could take their own time to play). This task also allowed us to collect data in the context of a demanding cognitive and reading process. A basic description of the articles' topic was given to the participants to help them complete the task. It can be found in Appendix A, Table A.3. When the task started, the first round's source article was already loaded (as shown by Figure 4.7). To play the second round, we asked participants to reach the source article by themselves from the Wikipedia search field.

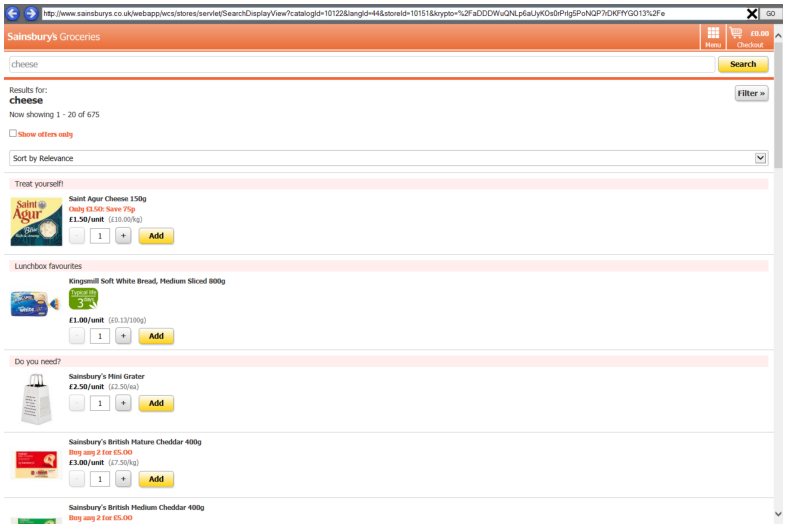


Figure 4.6: Shopping task.

4.3.2 Study Protocol

The study followed the following protocol.

1. At arrival, the participant was given the consent form about the study, with basic explanations about it.

2. Before starting a task, the participant was explained what was expected from her, and appropriate documentation was given (the list of the questions for the search task (cf. Appendix A, Table A.1), mock-up personal data for the shopping task (cf. Appendix A, Table A.2), the articles' topic main description for the game task (cf. Appendix A, Table A.3)). The order of the tasks performed by the participants were incomplete counter-balanced (Latin square).

3. Each task started with the eye tracker 9-point calibration (22-pixel radius), followed by a check test of the calibration (coloured squares to gaze at, display of the gaze location) cf. Chapter 3.1 for details.

4. Each task ended with a drift measurement (same format as the calibration).

5. At the end of the data collection, the participant was given a questionnaire for demographics, for judging her experience with tablets and with eye tracking, and for evaluating the browser she was using during the data collection. The questionnaire sample is provided in Appendix A, Figure A.1.

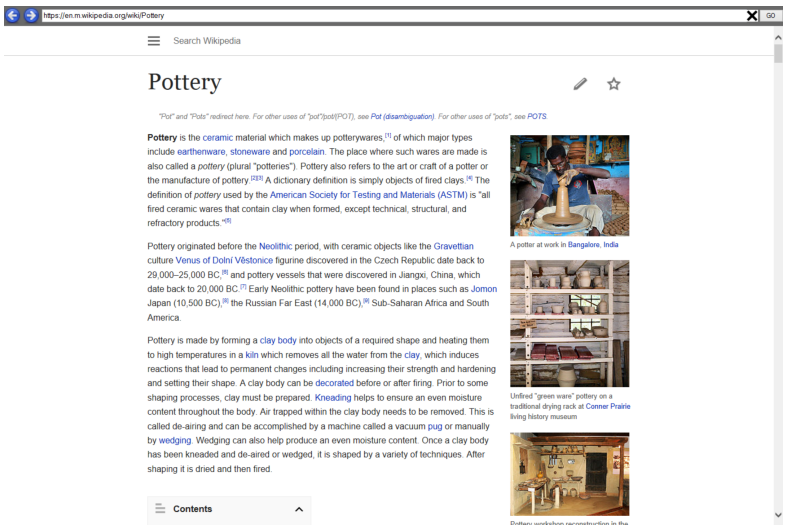


Figure 4.7: Game task.

4.3.3 Study Architecture and Data Collection

We favoured tablets for the data collection over other touch devices because of their reasonable size, prevalence, and compatibility with eye trackers. We used a Microsoft Surface Pro 3 (2160×1440 pixels resolution). We chose the Tobii X2-60 eye tracker (60 Hz), designed for studies on smaller devices. We selected the Tobii's rack after running a pilot study to validate the rack would not impair the participant's interaction (cf. Section 4.2). We thought of including the Tobii rack in our apparatus for its compatibility with the eye tracker and its design: two guide bars prevent the users from placing their hands above the eye tracker, which prevents hand occlusion. Figure 4.8 shows how the stand and the eye tracker were set for the data collection. Participants sat on a chair to interact with the tablet.

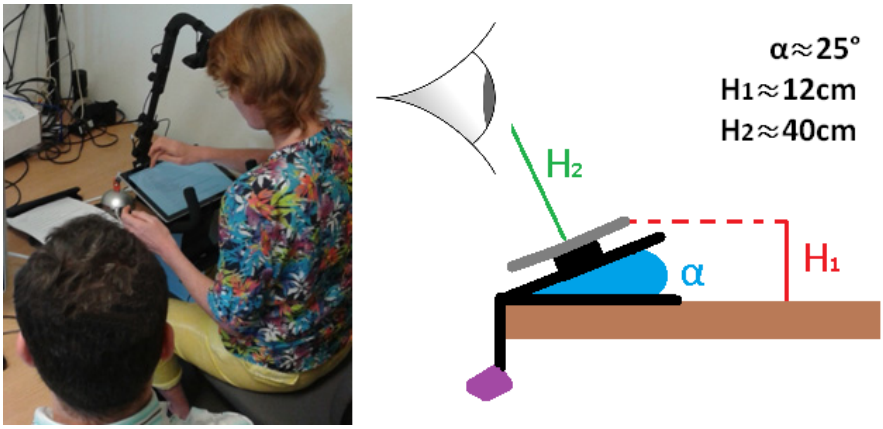


Figure 4.8: Eye tracker and stand configuration.

To keep the naturalness of the interaction, we did not ask the participants to limit their hand actions to the dominant hand: they were free to interact the way they liked.

The data collection consisted of retrieving the following information: touch input, on-screen gaze position, and tapped object characteristics. Touch data was collected through different steps. Details are provided in Section 3.2. The resulting files of the hand gestures contained the following information:

- timestamp,
- type of gesture (`TOUCH_EVENT`, `DRAG_EVENT` or `ZOOM_EVENT`),
- normalised position of the event,
- optional parameter, only set for `ZOOM_EVENT` to indicate the zoom factor.

We wrote an application (in C#) that retrieved gaze data samples from the eye tracker, and logged them into a file, which contained the following:

- timestamp,
- for each eye: estimated point of gaze in normalised screen coordinates
- for each eye: validity code (Tobii specifications to inform about the reliability of the tracking).

This application also ran the eye tracker calibration before each task and the drift evaluation afterwards. Fixations were computed post-hoc with OGAMA[2] on a spatial detection threshold of 22 pixels ($\sim 0.56°$ of visual angle). For this data collection, we implemented a web browser (C# WinForms application, providing a `WebBrowser` object that ran the Internet Explorer 11's engine). We decided to develop our own browser in order to easily

[2] `http://www.ogama.net/` (last accessed Jan. 2020)

get feedback from it and to offer a basic and sleek user interface for all participants. The browser had a dimension of 1440×960 pixels[3], with a viewport of 1440×914 pixels, topped by a navigation bar (illustrated by Figure 4.9). Participants evaluated the browser after the study and scored it 3.6±0.9 on average on a 5-point Likert scale. We report here few comments given by the participants regarding the browser: "*Difficulty in getting the keyboard out. Not very sensitive to touch*" (participant 5), "*Generally worked well but it sometimes dropped the keyboard when I was typing to select places to type. The search bar was also difficult to use - small to touch easily and it was hard to highlight specific text.*" (participant 6), "*I couldn't tell the difference to other browser.*" (participant 10).

Figure 4.9: Browser's navigation bar (truncated).

Participants evaluated the overall setup after performing the tasks, and gave an average of 3.6±0.8 on a 5-point Likert scale. Here are some of the participants comments regarding the overall setup: "*Position of hands not very comfortable. When I use tablets I keep them more vertical than horizontal.*" (participant 5), "*Most things were intuitive but with the exceptions from the previous section* [about the browser]", "It's probably not as natural as holding the device, but the rig did not substantially affect my experience." (participant 10), "*Simple, unobtrusive and straightforward to use.*" (participant 16). We categorised the tapped objects from their 3 distinct natures: HTML, browser or keyboard[4] elements. HTML and browser targets were tracked via the browser (using JavaScript injected code when the `OnWebBrowserDoClick` callback method of the `WebBrowser` object was called, as shown by the code preview in Appendix A, Figure A.2), the following relevant elements were written into a log file containing:

- timestamp,
- type of event:

  `AddrTextBoxTouched` when the browser's address bar was tapped on,

  `BackButton` when the browser's back button was tapped on,

  `click` when a webpage HTML element was tapped on,

  `GoButton` when the browser's go button was tapped on,

[3]This is the dimension of the full screen not in the high DPI mode on the tablet.

[4]In the thesis, *keyboard* should be understood at the *virtual* keyboard displayed on the tablet's touchscreen.

- target's size,
- target's relative position in the viewport,
- for HTML targets: the tag name (for instance `A`, `INPUT`, etc),

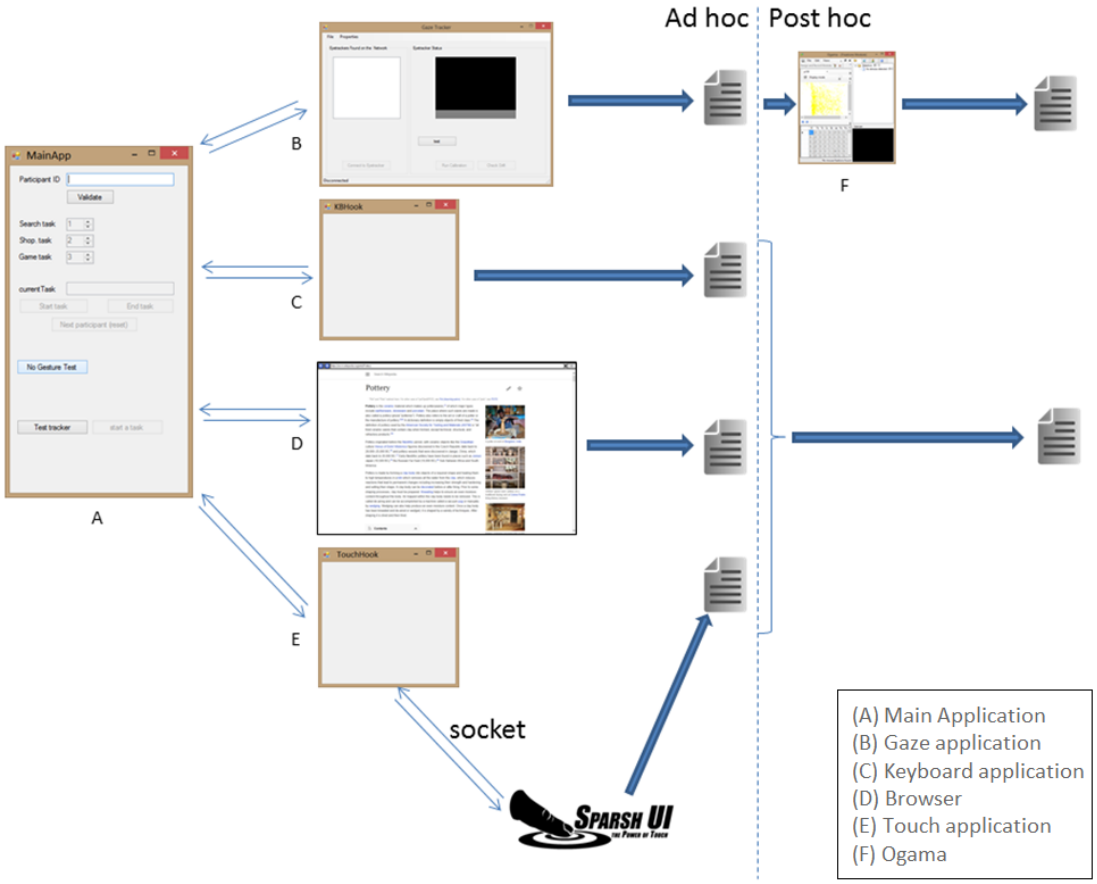


Figure 4.10: System architecture.

We made a specific application (in C#) to track the keyboard targets, with the help of the native Windows library Microsoft Keyboard Input API. The resulting log files contained:

- timestamp,
- key code.

In a post-hoc step, we then aggregated these different log files with the taps log files into a single file, based on the timestamps. This final file communicated the following information:

- timestamp,
- participant ID,
- task,

- tapped target type (`html`, `KB` or `browser`),
- tapped target name (in the case of type `html` it was the HTML element tag name (i.e. `A` or `INPUT`), in the case of the type `KB` it was the key code (i.e. `Return`, `Back` or `OemPeriod`), in the case of `browser` it was the keyword indicated which browser object was touched (i.e. `AddrTextBoxTouched`)),
- in the case of type `html` or `browser`: the tapped target abscissa; in the case of type `KB`: the literal value of the key (i.e. "`Return`", "`Back`" or "`.`"),
- in the case of type `html` or `browser`: the tapped target ordinate,
- in the case of type `html` or `browser`: the tapped target width,
- in the case of type `html` or `browser`: the tapped target height,
- the tap abscissa,
- the tap ordinate.

The flow chart in Figure 4.10 summarises the different steps aforementioned.

4.3.4 Data Collection Content Overview

We collected in total 574 675 touch data samples and 1 869 705 gaze data samples, from 24 participants (9 female, age 31.4±11). Flyers about the study (illustrated in Appendix A, Figure A.3) had been put in Lancaster University campus and in town, as well as handed over directly to people passing-by on the central square of the campus. In order to avoid people being conscious of their eyes being trackers, the study was presented as an Internet activity study, with a data collection on touch and behaviour. We did not pay the participants for the study, but offered refreshments and snacks instead. The study lasted about an hour, depending on the speed of the participants. We did not exclude participants before evaluating how the eye tracker worked with their eyes (impossibility to calibrate, cf. Chapter 3.1.4 for details). Being a highly international environment, only few participants, nine of them, were English native speakers. English understanding was required to complete the tasks, and the self evaluation of English level among participants scores 4±1.5 on a 5-point Likert scale. A participant commented "*For students who English is not their first language, these tests are a little bit difficult.*" (participant 3). All of them were familiar with Internet browsing and they gauged themselves as experienced with tablets and touch devices (respectively 3.3±1.2 and 3.7±1.2 of average on a 5-point Likert scale). Some participants needed visual correction during the study (7 wore glasses, 3 wore contact lenses). All but one were right-handed. We tracked three types of touch

gesture actions: taps (representing 72 % of the data set), pans (representing 28 % of the data set) and zooms (representing <1 % of the dataset). These statistics reveal that the tablet was large enough for the participants to interact with it comfortably (almost no zooms). The tapped object distribution is as follows: keyboard (68.9 %), HTML (26.5 %), browser (4.6 %).

4.4 Fixations around Tap Moment

In this section, we present the general characteristics of the fixations which occurred at tap, the study unit in this chapter. Fixations are one of the key metrics in eye tracking studies as they reflect, in some aspects, the attention given by the subjects in their activity.

4.4.1 Spatial Distribution of All Fixations at Tap

In this section, we give a first estimation of the gaze behaviour during tap from a very coarse approach. The metric we report describes a spatial feature of the fixations: the distance between the fixation position on the tablet and the tap position on the tablet. Figure 4.11 illustrates the spatial distribution of *all* the fixations occurring in a time window of 2 seconds around the tap moment. The position of the fixations on the plot is relative to the tap location.

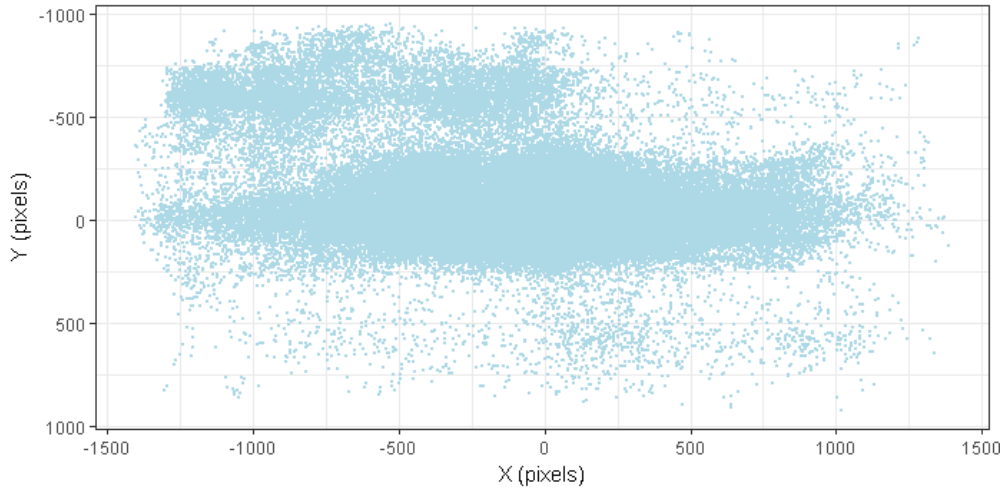


Figure 4.11: Spatial distribution of the fixations relative to the tap position (2 seconds around the tap).

From Figure 4.11, there is an indication that fixations are scattered around the tap position, but two clusters appear: the first one is concentrated close to the tap point

(centre of the graph), while the second is located at the top left side of the tap position. On Figure 4.12, we illustrate the distances for which a certain percentage of the fixations are under that distance. For example, half of the fixations observed in the 2 seconds window are contained within 268 pixels and 75 % are contained within 594 pixels.

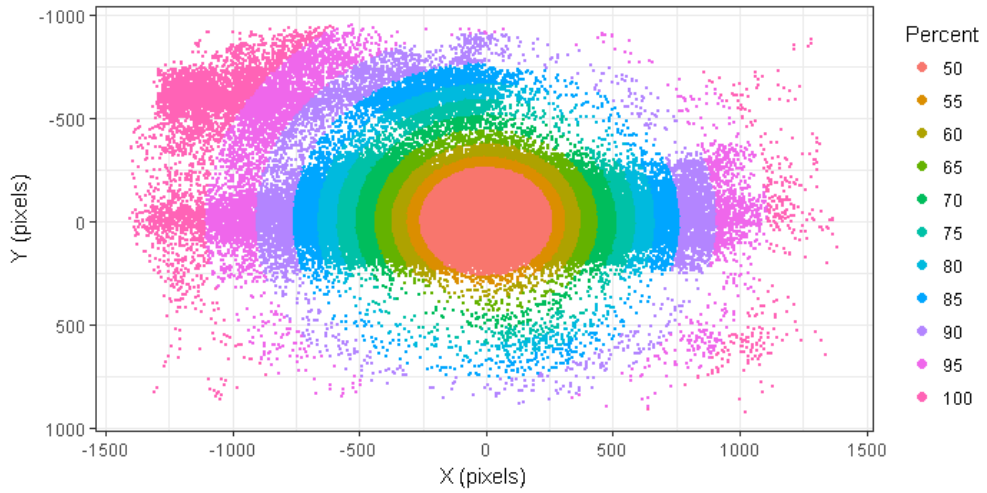


Figure 4.12: Spatial distribution of the fixations relative to the tap position (2 seconds around the tap).

Based on the representation mentioned above, we question the role played by time on the dispersion of the fixations. We therefore investigate whether there is a narrower time window (within the centred 2 seconds time window previously chosen) within which the dispersion of the fixations would be minimised (meaning that for a given percentage, the distance from the tap position obtained with another time window would be shorter than the one found with a centred 2 seconds time window). Figure 4.13 shows how the distances varies at different *non overlapping* time windows. It clearly indicates that for a certain time window (here found to be -0.300 s to -0.250 s), the distribution of fixations is much closer to the tap point: half of the fixations observed in this window are contained within 69 pixels and 75 % are contained within 133 pixels. We conclude, therefore, that at some moment in time *before* tapping (which value should have been somewhere in or near the non sliding time window mentioned above), gaze is approaching at the closest to the target. Further findings aligned with this preliminary presentation are detailed in Section 4.4.3.

4.4.2 Number of Fixation Before and After Tap

We find that on average, there are 4 fixations before and 3 fixations after the tap moment within a 2 seconds window centred on the tap moment (before: 3.72±1.75 fixations, after:

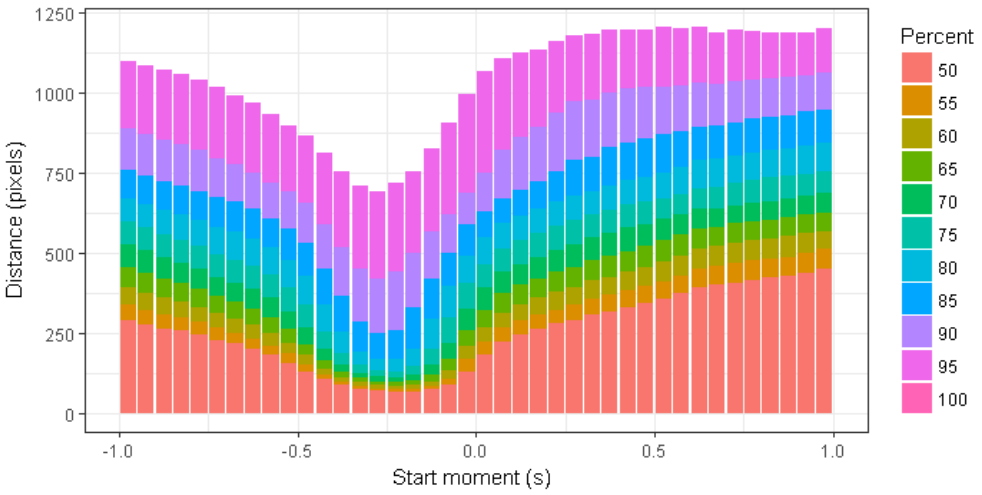
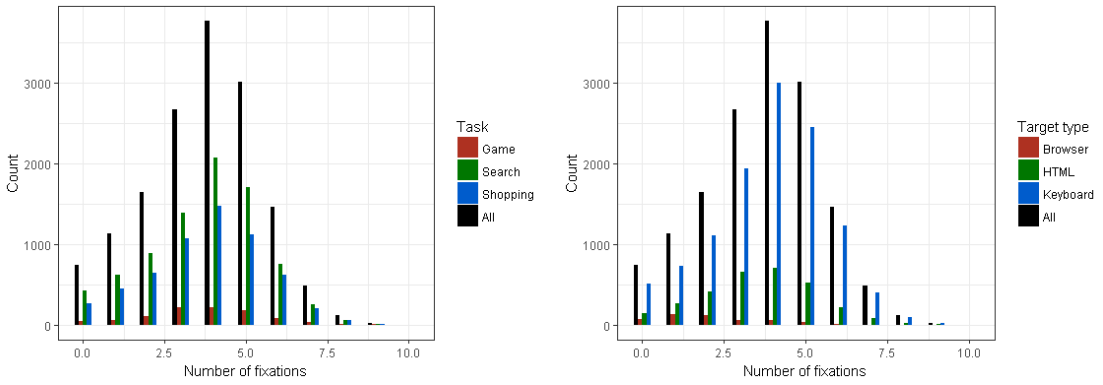


Figure 4.13: Distance percentages of the fixations relative to the tap position at different time windows relative to the tap moment (50 ms wide).

2.93±1.65 fixations). Figures 4.14 and 4.15 respectively show the number of fixations strictly before and after (or at) the tap, per task and per target type.

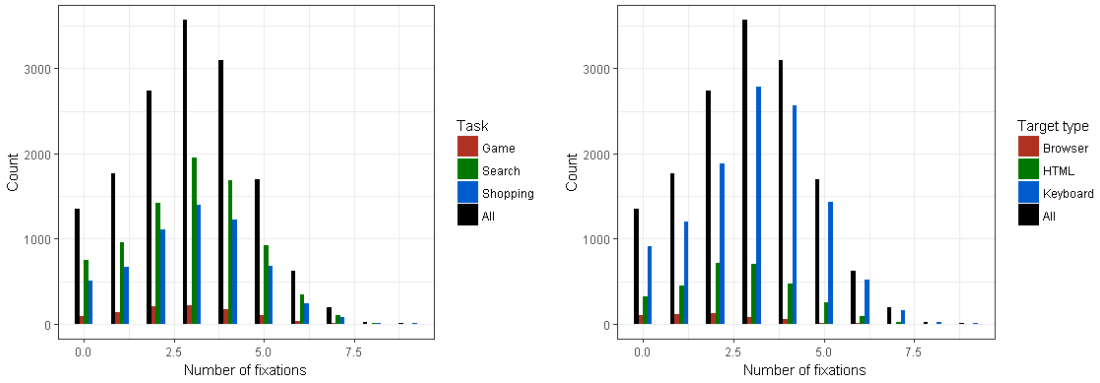
From Figures 4.14 and 4.15, we can observe that in most cases, the distributions of the number of fixations before or after the taps in the different conditions are the same as for the general case. A main exception is found for the target of type "browser" (elements of the browser UI). On average, for this type of target, the participants perform 2 fixations before and after tapping, which is probably due to the static and known positions of these elements, resulting in reducing the visual search prior tapping. The fewer number of fixations after tapping a browser element may be explained by the navigation activity: after tapping an element of the browser, the participant need to wait for the response triggered by that element.



(a) Per task.

(b) Per target type.

Figure 4.14: Number of fixations strictly before the tap moment.



(a) Per task.

(b) Per target type.

Figure 4.15: Number of fixations at and after the tap moment.

4.4.3 Relationship between Spatial and Temporal Distribution of the Fixations around Tap

In Section 4.4.1, we previously showed that most fixations are at a closer distance from the tap point *before* the tap happens. In this section, we look for a better estimation of the moment when fixations are at the closest to the tap position. To do so, we estimate the relationship between the fixations' starting moment and their distance to the tap position. We choose a Generalised Additive Model (cubic spline) for the estimation method because of the initial assumption of nonlinear model (which is strongly suggested by the bar chart in Figure 4.13) and the fact that we do not expect the data to be normally distributed (gaze is always "on", so fixations happen all the time and it seems illogical to think that the starting moment of the fixations would be normally distributed in time). The resulting estimation model is plotted in Figure 4.16.

The estimation plot validates the two observations made in Section 4.4.1: (1) in time, the fixations tend to approach the tap point before the tap occurs, and then recede from the tap position again (typical "V-shaped" curve which appeared also in Figure 4.13), and (2) the approximation we made earlier was not far from the time we estimate the distance to be the closest to the tap point (by finding **the minimum** of the estimation curve). We find that 0.338 seconds before the tap occurs, gaze approaches the tap point within a distance of 159 pixels.

Looking at the relationship between the fixations' temporal and spatial characteristics was inspired by the work done in studies dealing with the correlation between gaze and mouse (i.e. [15, 82, 97, 123, 174, 217]). However, the major difference between these existing works and the work presented in this thesis is the nature of the "manual input":

mouse is a continuous input whereas tap is a punctual input. Nevertheless, our findings are still coherent with the results given about the correlation between gaze and mouse: for instance, Liebling and Dumais [123] reported a lag between gaze and mouse of -250 ms to -100 ms, by about 74 pixels.

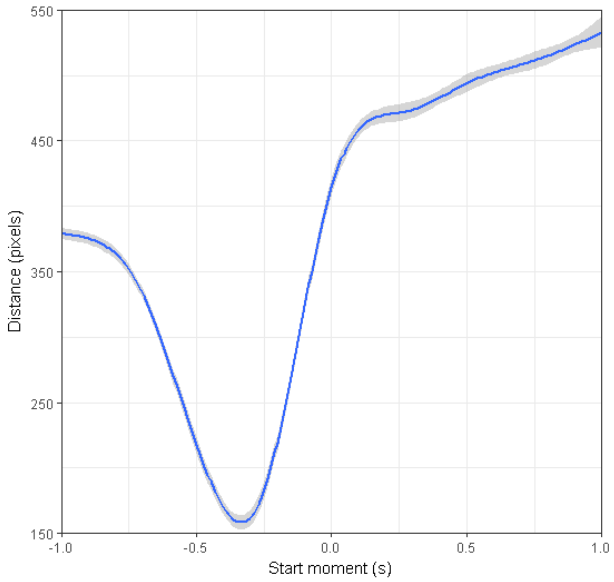


Figure 4.16: Fixation start moment vs. fixation distance (relative to tap moment/position).

4.4.4 Impact of Individuals, Tasks and Target Types

So far, we have presented a coarse estimation of the relationship between gaze and tap for *all* participants, tasks and target types together. We want to investigate how these factors influence the relationship between gaze and tap.

For each participant, task and target type, we estimate the relationship between the fixations' spatial and temporal characteristics relatively to taps with the same model given in Section 4.4.3. In order to compare them, we report the descriptive statistics related to the **minima** given by the estimation model.

4.4.4.1 Across Participants

Estimations for each participant are plotted in Figure 4.17, and the minima's value reported in Table 4.4. For all participants, gaze leads the tap. However, disparity is observed between each participant. This difference is more important in space (Δ=181 pixels, σ=46.1 pixels) than in time (Δ=230 ms, σ=47 ms). We deduce that the distance within which participants keep their gaze away from the targets is a personal feature

(some participants as we observed during the data collection tend to "follow" their gaze with the finger, while others would keep the finger still will skimming the display and only move to tap when they need), and that it influences the correlation between gaze and tap.

Table 4.4: Fixation start moment vs. fixation distance (minima, per participant).

P[a]	S.M.[b]	Dist.[c]	P[a]	S.M.[b]	Dist.[c]	P[a]	S.M.[b]	Dist.[c]
#1	-371	87.8	#9	-384	200.3	#17	-356	176
#2	-396	89.5	#10	-305	170.8	#18	-204	151.2
#3	-339	170.1	#11	-373	163.6	#19	-325	133.1
#4	-325	162.3	#12	-315	221.9	#20	-287	66.4
#5	-350	143.7	#13	-333	85	#21	-333	132.4
#6	-369	178.8	#14	-340	114.2	#22	-415	158.9
#7	-297	178.4	#15	-357	220.1	#23	-372	177.1
#8	-317	184.1	#16	-305	200.7	#24	-433	247.7

[a]Participant [b]Start moment (ms) [c]Distance (pixels)

4.4.4.2 Across Tasks

Estimations for each task are plotted in Figure 4.18 and the minima's value reported in Table 4.5 (left part). Again, we observe that gaze precedes touch, and that the tasks influence the spatial dimension (Δ=31.9 pixels, σ=17.3 pixels) rather than the temporal dimension (Δ=22 ms, σ=11 ms). For the search task, the minimum of the estimation graph has a greater Y-axis value than for the other two tasks. We can interpret this as a consequence of SERPs being systematically queried by the participants for that task: users mainly follow the first link after scanning a few, confirming a common behaviour reported in [107]. Thus, taps were possibly already "prepared" to be performed while participants were still scanning the webpage, with the consequence that they did not need to acquire the target again when tapping afterwards. The task nature has therefore a clear influence on the correlation.

4.4.4.3 Across Target Types

Estimations for each target type are plotted in Figure 4.19 and the minima's value reported in Table 4.5 (right part). Gaze still precedes touch, and the temporal difference

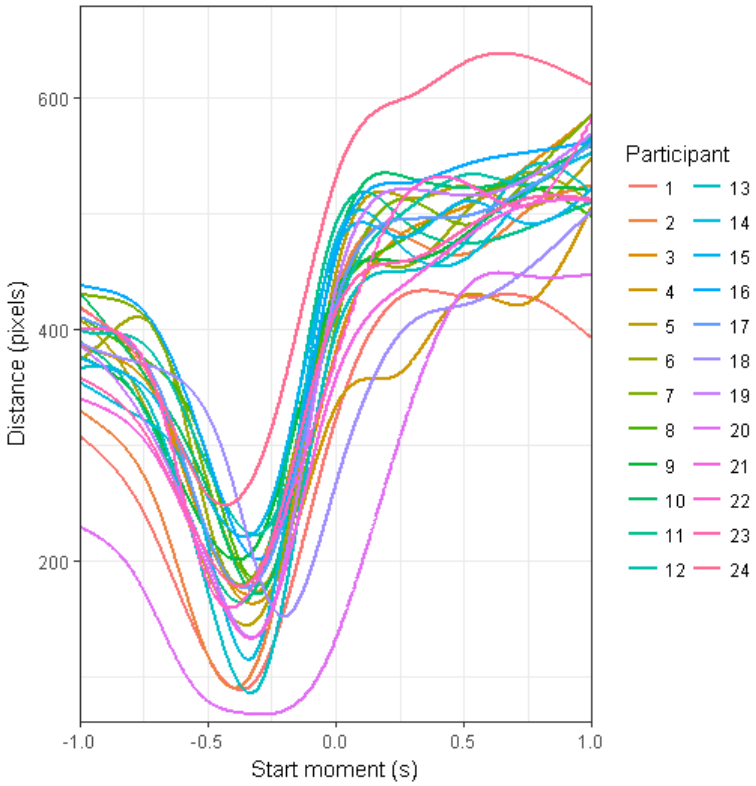


Figure 4.17: Fixation start moment vs. fixation distance (relative to tap moment/position, per participant).

varies less than in space (respectively Δ=48 ms, σ=25 ms; Δ=32.4 pixels, σ=16.2 pixels). Fixations for the keyboard objects seem to happen earlier before the touch, with a farther distance to the target. We explain this by the potential learning effect in typing and using the browser. Participants took less time "searching" the target, and did not need to visually focus on it as they mentally already knew where it was. Thus, the tapped target plays a role in the correlation, depending on its likelihood to be known in advance.

Table 4.5: Fixation start moment vs. fixation distance (minima, per task/target type)

Task	S.M.$^{(a)}$	Dist.$^{(b)}$	Target type	S.M.$^{(a)}$	Dist.$^{(b)}$
Search	-328	174.6	Keyboard	-325	167.7
Shopping	-338	142.7.9	HTML	-363	135.3
Game	-350	147.2	Browser	-373	152.7

$^{(a)}$Start moment (ms) $^{(b)}$Distance (pixels)

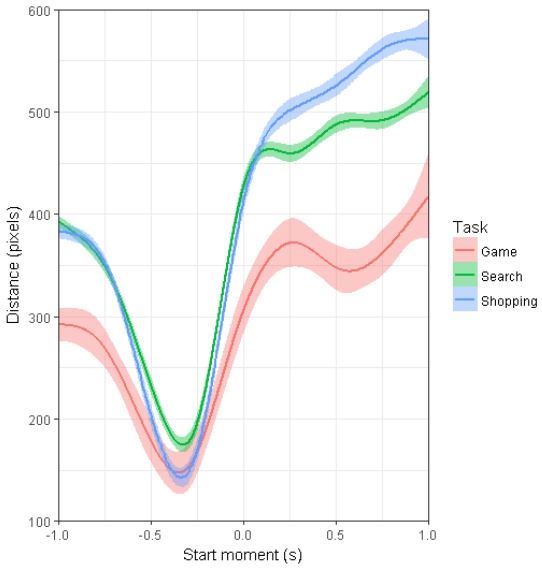


Figure 4.18: Fixation start moment vs. fixation distance (relative to tap moment/position, per task).

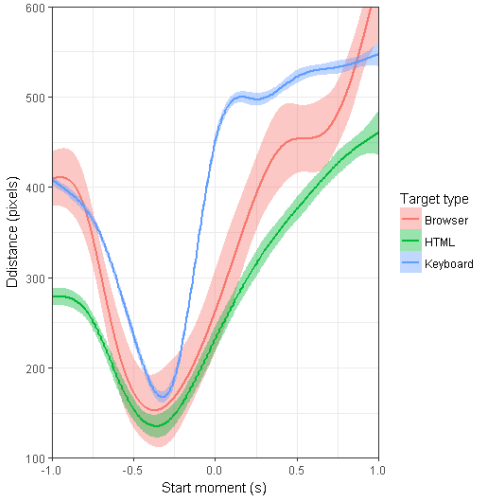


Figure 4.19: Fixation start moment vs. fixation distance (relative to tap moment/position, per target type).

4.5 A Specific Fixation: $F_{Closest}$

This section explores the characteristics of a specific fixation, that we labelled $F_{Closest}$. This specific fixation is the one which arises *before* a tap happens *at the closest* to this tap's position. Describing $F_{Closest}$ confronts the estimations given earlier.

4.5.1 $F_{Closest}$ General Characteristics

We evaluate the characteristics of $F_{Closest}$ in the manner that we presented fixations in Section 4.4: the start moment of the fixation (relatively to the tap moment) and the distance between the fixation and the tap position on the display.

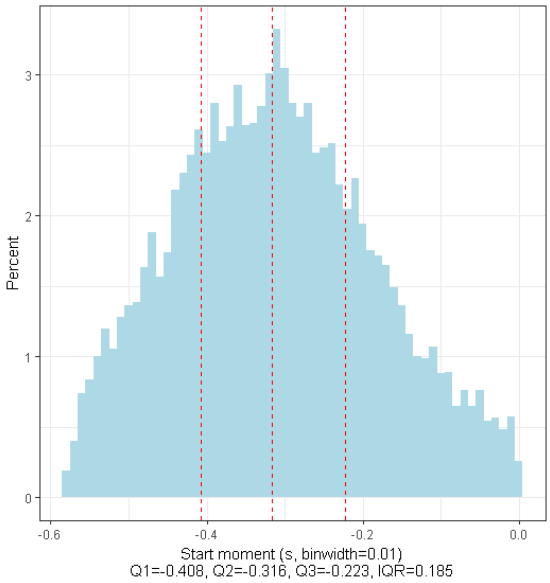


Figure 4.20: $F_{Closest}$'s start moment histogram and quartiles.

For each tap, we retrieve $F_{Closest}$ in a window starting 0.6 s before the tap moment. This boundary has been chosen based on the minimum point for all fixations reported in Section 4.4.3 and the difference within participants in 4.4.4.1 (-0.338 − 0.230 ≈ -0.6).

Figures 4.20 and 4.21 respectively show the histograms of $F_{Closest}$'s start moment (to the tap moment) and distance (to the tap position), and the associated quartiles values. The statistical modes of the start moment and of the distance are respectively -0.341 s and 32.8 pixels. The figures indicate that $F_{Closest}$ is normally distributed in time, whereas spatially it is positively skewed.

4.5.2 Spatial Distribution of $F_{Closest}$ Relative to Taps

We study the spatial distribution of $F_{Closest}$ around the tap points, plotted in Figure 4.22. The standard deviation on the X-axis is 148.3 pixels, and 87.7 pixels on the Y-axis. Both the mean position (-24.4 pixels,-28.8 pixels) and the median position (-8.7 pixels,-12.7 pixels) show an offset which can be explained: web users tend to look more at the top-left part of the page [144]. They are also observed on Figure 4.22.

Figure 4.23 and 4.24 represent respectively $F_{Closest}$'s mean and median positions, for each participant. Although they are spread around the global mean and median positions, they remain generally offset towards the top-left direction, most notably for the mean positions: 20 participants (83 %), against 13 for the median position (54 %) - we regard the participants whose median/mean positions show up in the top left quarter of the graph.

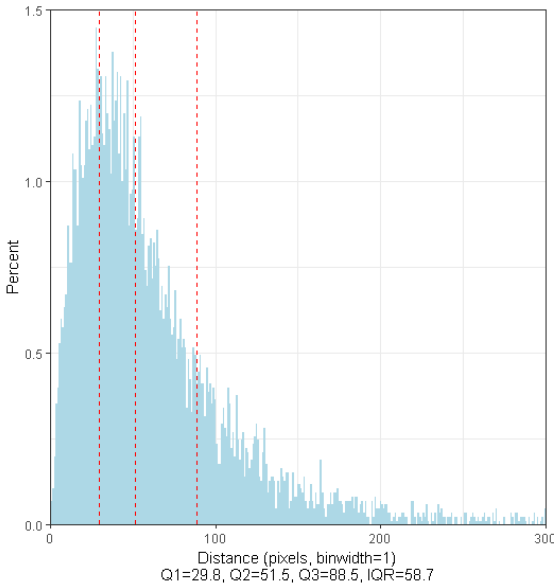


Figure 4.21: $F_{Closest}$'s distance histogram and quartiles.

Figure 4.25 illustrates $F_{Closest}$'s mean and median positions for each task. For both search and shopping tasks, we notice the mean and median positions do not vary more than 8 pixels around the overall values in each direction. For the game task, we observe that $F_{Closest}$'s mean and median positions are closer to the tap point. In this task, the targets' position (mostly links) could not be "learnt" nor "anticipated" by the participants, contrary to the other tasks. For the search task, we suppose that learning effect of selecting the first link(s) in the SERP, as well as a tap anticipation (as discussed in Section 4.4.4.2) appeared from the participants. For the shopping task, learning effect may have come from the commercial website interaction. We suppose the learning effect and the tap anticipation brought by a task can influence the distance between gaze and tap: when there is none of these effects, gaze acquires targets with a closer distance.

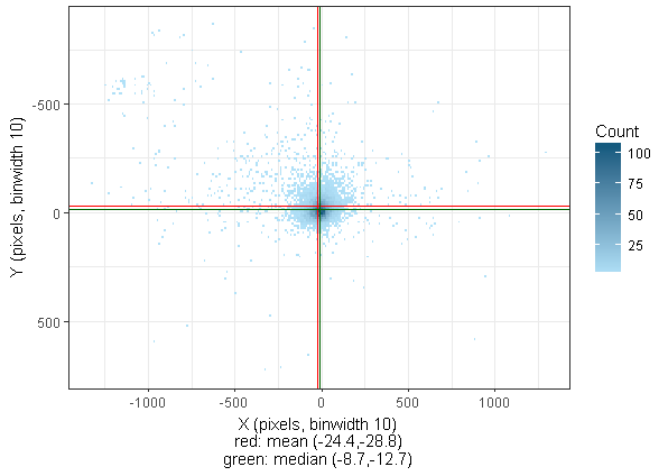


Figure 4.22: $F_{Closest}$'s mean and median positions around tap position.

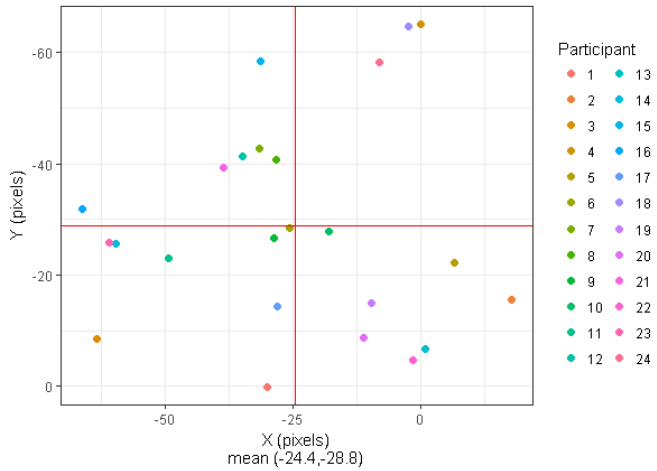


Figure 4.23: $F_{Closest}$'s mean position around tap position (per participant).

$F_{Closest}$'s mean and median positions for each tapped object type are represented in Figure 4.26. There is an expected difference for the browser elements. Being situated in the top part of the display, $F_{Closest}$'s mean and median positions are not likely to show an offset on the top-left side of the screen. Browser elements allowed navigation and triggered changes in the viewport situated below, hence an opposite offset direction. The case of the HTML elements shows a very small vertical offset (less than 7 pixels for both mean and median values) indicating that gaze is more often vertically aligned with the targets. We can interpret this result as an effect of reading: most HTML targets were links ($\approx 40\%$) and text input fields ($\approx 26\%$).

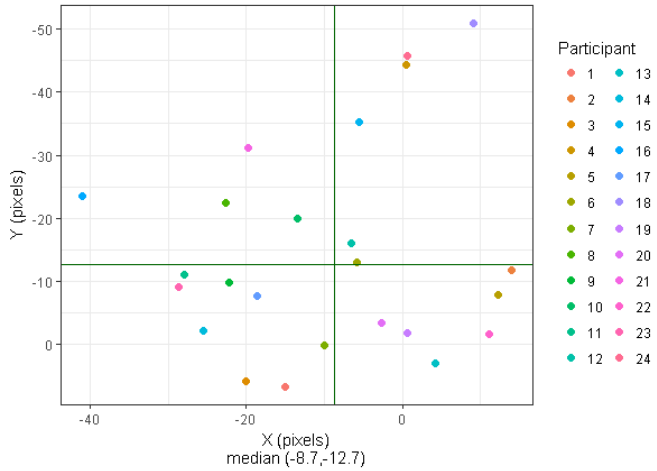


Figure 4.24: $F_{Closest}$'s median position around tap position (per participant).

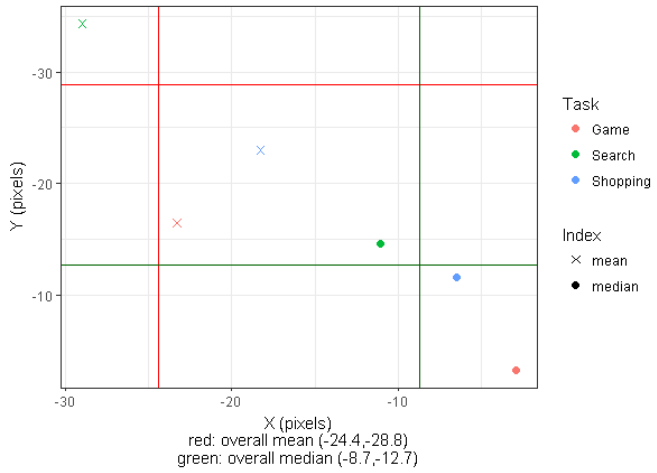


Figure 4.25: $F_{Closest}$'s mean and median positions around tap position (per task).

4.5.3 Relationship between $F_{Closest}$ and Tap

In the previous sections, we described $F_{Closest}$ relatively to the taps without considering the position nor the size of the target objects. We want to find out whether the distance between $F_{Closest}$ and the tap vary depending on the target's position and size. To do so, we compute the Pearson correlation between the horizontal (respectively vertical) distance of $F_{Closest}$ with the tap's position and either the abscissa (respectively the ordinate) or the width (respectively the height) of the target. However, we only run the Pearson correlation test on a subset of the data. Keyboard targets should be discarded due to the potential differences between the typing mechanisms and HTML/browser targets selection. HTML elements that are containers have also been ignored: firstly because these elements have often been tapped on by mistake (for instance when missing a link or a checkbox), secondly because container elements, by definition, are not meant

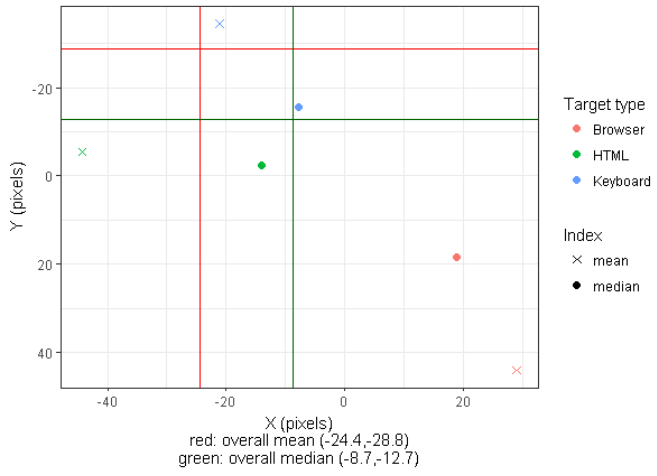


Figure 4.26: $F_{Closest}$'s mean and median positions around tap position (per target type).

to be tapped and sometimes covered a large portion of the webpage, that makes the correlation with tap totally irrelevant. These container elements are `DIV`, `BODY`, `TD`, `FIELDSET`, `SPAN` and `HTML`. The Pearson correlation is reported in Table 4.6, we do not find any relationship between the position or the size of the target with $F_{Closest}$'s distance to the tap.

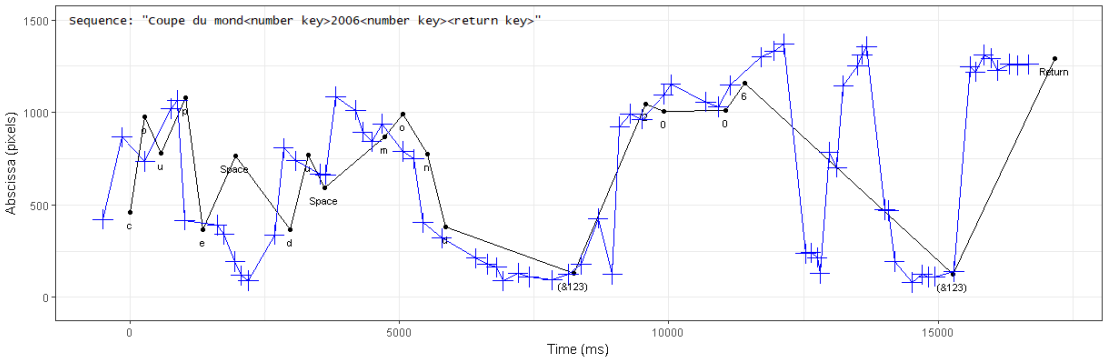
Table 4.6: Pearson correlation between the $F_{Closest}$/tap distance and the target position/size.

Target's characteristics	Distance between $F_{Closest}$ and the tap (relative)*	(absolute)*	(relative)**	(absolute)**
Position	0.04	-0.10	-0.04	-0.13
Size	-0.26	0.17	-0.02	0.02

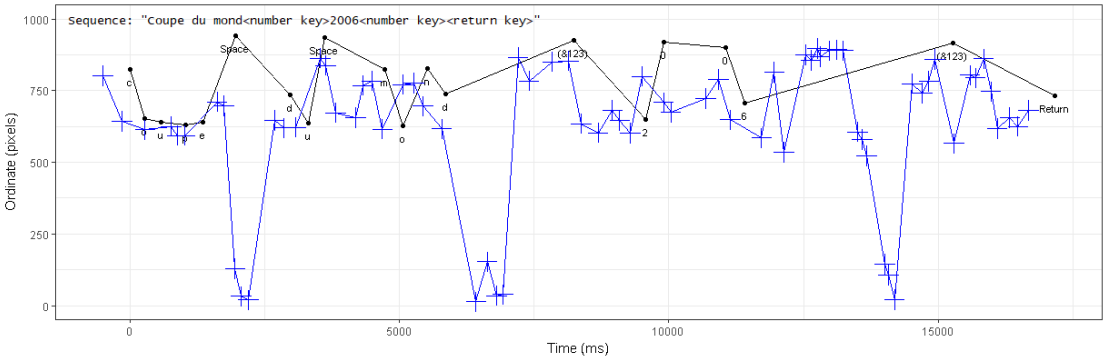
(*)horizontal dimension (**)vertical dimension

4.6 Typing

The keyboard typing events are constituting a significant part of the taps dataset we obtained from the data collection (cf. Section 4.3.4). Since they are the major group of targets, and because typing takes a specific place in computer interaction, we desire to know how strong gaze and typing are aligned.



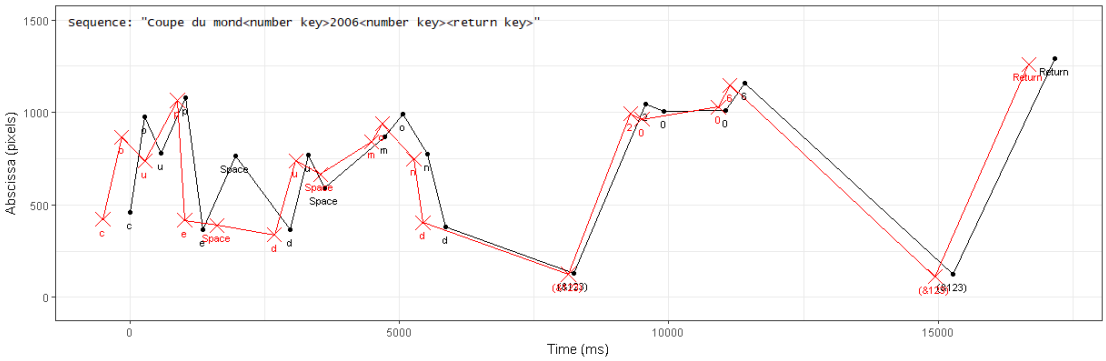
(a) Horizontal coordinates.



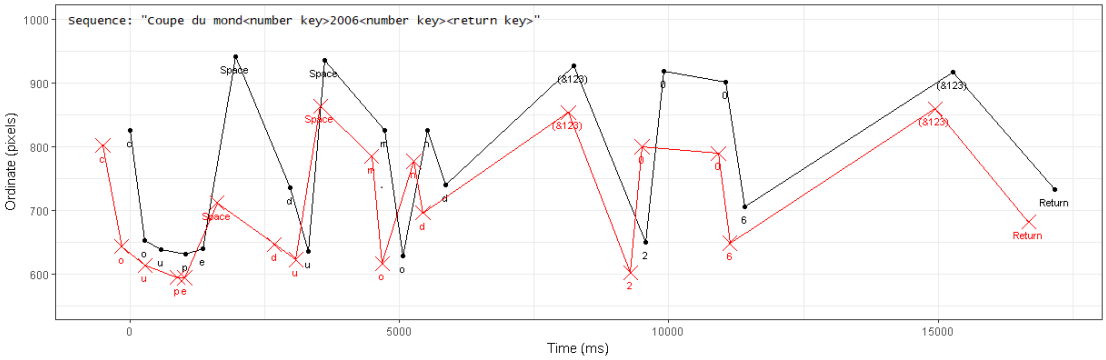
(b) Vertical coordinates.

Figure 4.27: Fixations (blue) and taps (black) during a typing sequence.

To do so, we first compare the fixations' positions and moments to the taps' positions and moments during a typing sequence. Figure 4.27 illustrates an example of a typing sequence performed by a participant, in each dimension of the tablet's screen (horizontally and vertically). For both graphs 4.27a and 4.27b, the X-axis is the time relatively to the first tap of the typing sequence, and the Y-axis is, respectively, the horizontal and the vertical coordinates of the position on the display. The black line connects the taps of the typing sequence (black dots on the plot), whereas the blue line connects the fixations' start moment occurring during this same typing sequence (blue crosses on the plot, we select all the fixations between the value of $F_{Closest}$ corresponding to the first tap of the sequence and the value of $F_{Closest}$ corresponding to the last tap of the sequence).



(a) Horizontal coordinates.



(b) Vertical coordinates.

Figure 4.28: $F_{Closest}$ associated with taps (red) and taps (black) during a typing sequence.

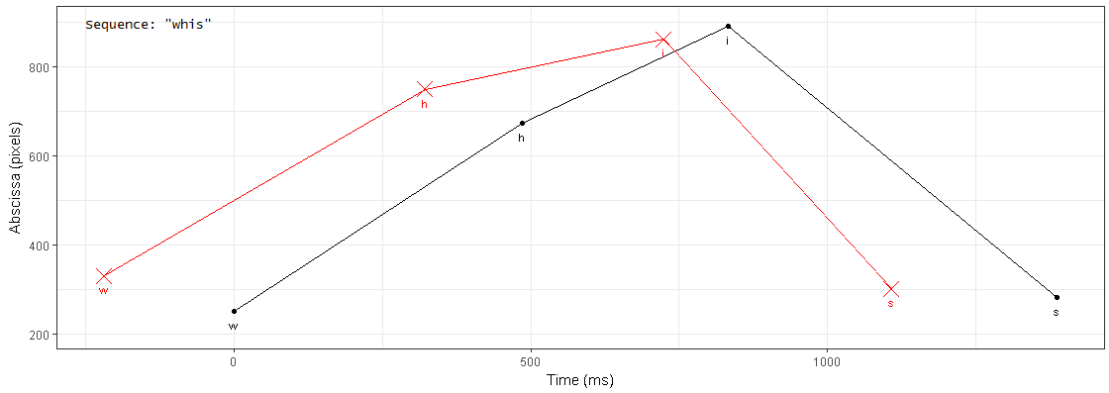
We observe from this example that gaze is loosely following the tapping flow in space. In time, an offset appears (as gaze precedes the touch input). However, there are cases where the fixations seem to completely "leave" the tapping pattern. This is a typical behaviour of when a participant was assessing her typing on the text input location of the screen, or simply when looking for the key's position through the keypad.

Because of the behaviour mentioned above, we want to check if only *some* relevant fixations are showing a better alignment with typing. The fixations we keep in the

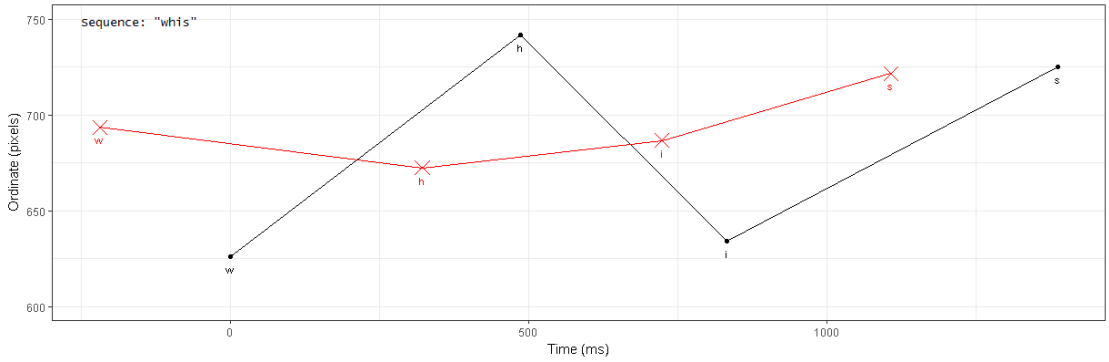
following subset are *only* the associated $F_{Closest}$ of each tap making the typing process.

Figures 4.28 illustrates the same typing sequence example than in Figure 4.27, with the partial gaze representation (which only includes $F_{Closest}$) in red.

The similarity between the "path" connecting each $F_{Closest}$ and the "path" connecting each tap in the sequence is even more obvious than when considering *all* the fixations during the sequence. The temporal offset we notice on the graph is coherent with the description we made of $F_{Closest}$ in Section 4.5.



(a) Horizontal coordinates.



(b) Vertical coordinates.

Figure 4.29: $F_{Closest}$ associated with taps (red) and taps (black) during a typing sequence (poor alignment).

From direct observation of the participants' interaction during the data collection, we clearly noted that some participants tended to type without "closely" looking at the keyboard. Therefore, the alignment we find between $F_{Closest}$ and the taps during typing also varies a lot from one participant to another, since some of them showed an habit of continuously checking their input while typing, when others preferred a two steps approach: first focusing on typing the sequence and then looking at the text input location to verify their input, or elsewhere in the webpage to continue with their ongoing activity. Figure 4.29 shows an example of a poor alignment between $F_{Closest}$ and the taps

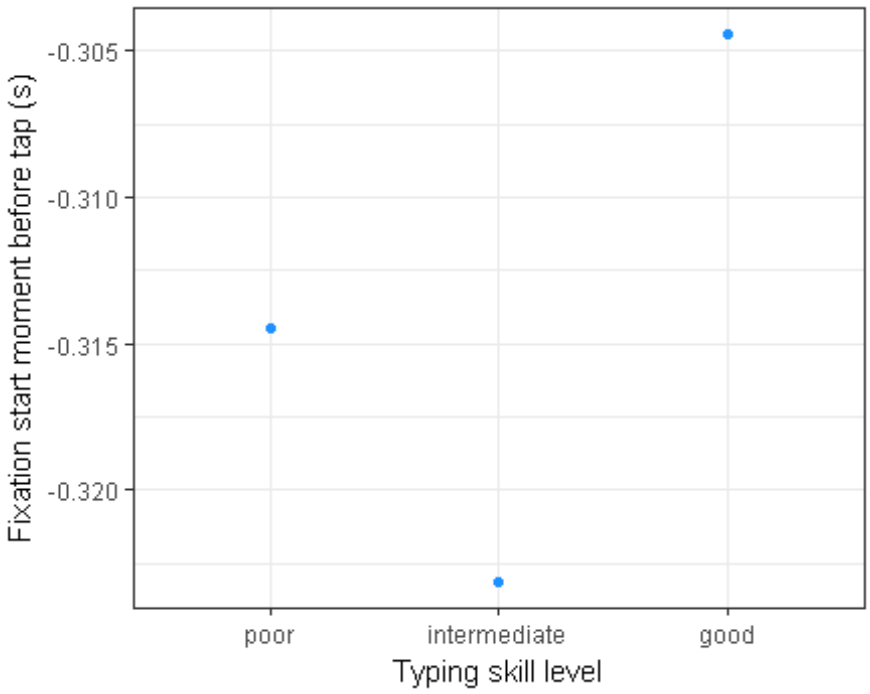


Figure 4.30: Average $F_{Closest}$ start moment before keyboard tap per typing skill level.

for a participant. In this example, gaze alignment is observed on the horizontal axis, but not as much on the vertical axis. The gaze strategy, for this particular participant, favours a left/right search on the keyboard rather than a top/down search.

From the video material we recorded during the data collection, we classify the typing skill level of the participants in 3 groups: poor, intermediate and good. The classification is done based on the observation of the speed, dexterity and accuracy of typing, and results in the following proportion of our population sample: 8 % (2 participants) showed a poor typing skill, 50 % (12 participants) showed an intermediate typing skill and 42 % (10 participants) showed a good typing skill. We investigated how the typing skill impacts the relationship between $F_{Closest}$ and the taps (on the keyboard).

Figure 4.30 shows the average time difference between key taps and their corresponding $F_{Closest}$ per typing skill level.

The average values found for each group are in the same range (around -0.315 s) but good typers seem to require less time between gaze acquisition and tap. We found a significant different between intermediate (-0.323 s) and good (-0.304 s) typers (one way ANOVA $F(2,7081) = 18.24$, $p < 0.01$; pairwise comparisons Tukey $p < 0.01$). Statistical tests may not find a difference between the poor typists and any other group due to their low representation in our population, and/or the temporal irregularity of their typing during typing sequences.

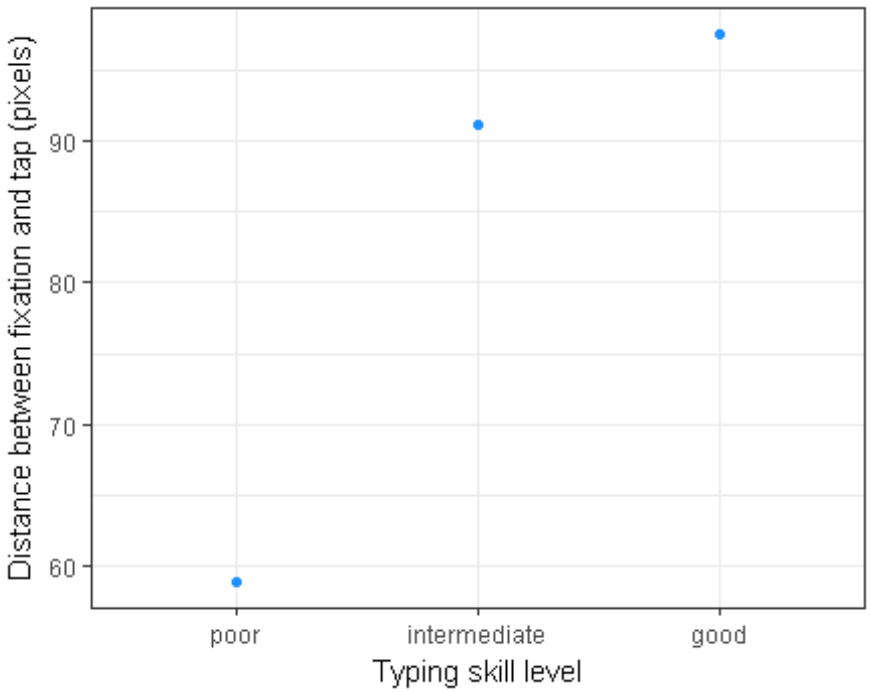


Figure 4.31: Average distance between $F_{Closest}$ and keyboard tap's locations per typing skill level.

We also investigated the spatial differences between groups. Figure 4.31 illustrates the impact of the typing skill level on the average distance between $F_{Closest}$ and the tap location (when typing).

We found that the distance depends on the typing skill level for all groups (Kruskal-Wallis[5] rank sum test $\chi^2 = 105.65$, df $= 2$, $p < 0.01$; pairwise comparisons Wilcox with Bonferroni adjustment all $p < 0.01$).

The average distance between $F_{Closest}$ and the taps for each dimension of the screen (X and Y) is respectively illustrated by Figures 4.32 and 4.33.

We found a statistical difference between each groups for the Y-axis, indicating that the distance between gaze ($F_{Closest}$) and tap increases with the typing skill level, mainly due to the vertical dimension (one way ANOVA $F(2,7081) = 60.17$, $p < 0.01$; pairwise comparisons Tukey all $p < 0.01$). This vertical difference can be explained by the fact that typers with a poorer skill tend to look at the keyboard more than the location where the typing takes effect on the tablet's interface.

[5]This test differs from the previous test because the distance between $F_{Closest}$ and taps is positively skewed (cf. Figure 4.21).

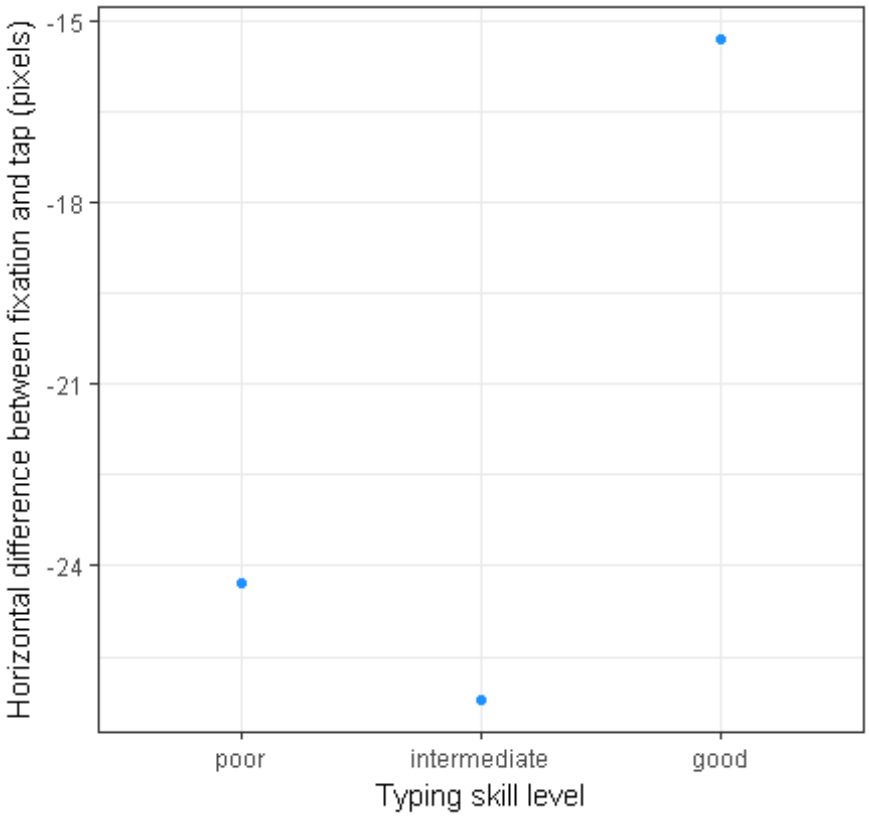


Figure 4.32: Average horizontal distance between $F_{Closest}$ and keyboard tap's locations per typing skill level.

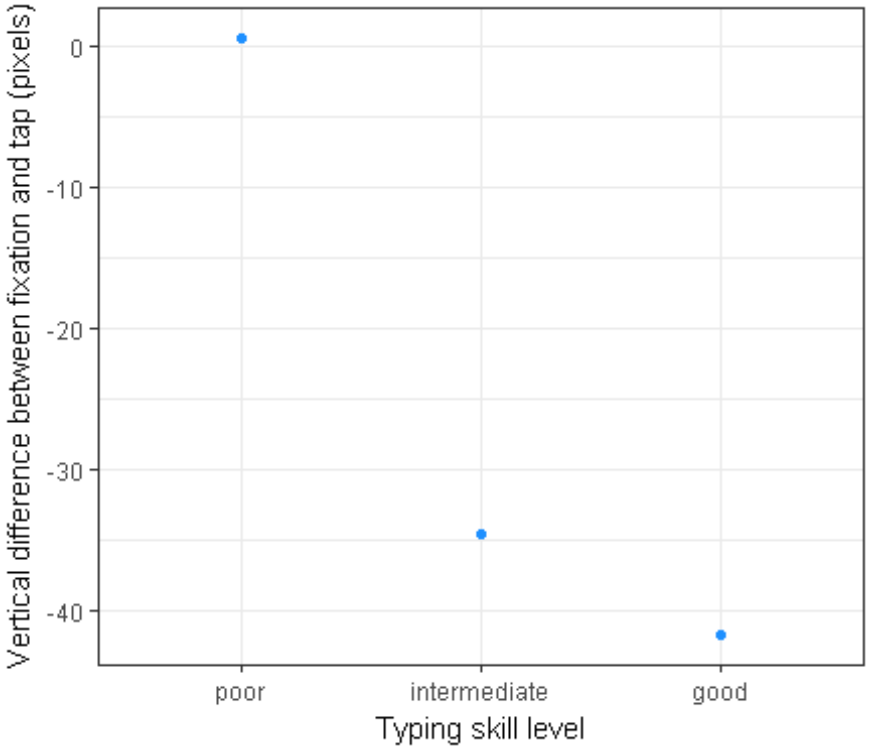


Figure 4.33: Average vertical distance between $F_{Closest}$ and keyboard tap's locations per typing skill level.

4.7 Discussion

Our results gave a coarse description of the correlation between tap and gaze in a natural context. Fundamental understanding of this correlation supplies an insight of how the gaze behaves along with hand activity, of which tapping is a critical part. Since further equipment to sense tapping is not required on tactile devices, it can be easily used to estimate where gaze was located just before the tap, and if an eye tracker where to be running, sense potential drift in the calibration during sessions.

Even finer results were limited by our study design. We did not consider taps in relation with the whole webpage content, and therefore cannot understand how elements other than the tapped one influenced the correlation. For instance, the salience of the webpage areas may affect the visual attention [187], and therefore the visual strategy chosen to skim the page. As we wanted to preserve the naturalness of the tasks in the study, we collected data from stimuli that varied across participants, and they tapped on different elements since they were free to browse at will to complete tasks. It was therefore a decision that led to a heterogeneous dataset. This diversity can partly explain the large variance observed in the statistical descriptions of the fixations we gave.

Potential applications from the results we found regarding the correlation between gaze and hand should exploit the expected distance and time between gaze and taps to estimate when a user is not able to locate a target adequately. If a first stage of personalised model is possible during interaction, an intelligent system can detect deviance between the expected personalised model and the actual measurement of the correlation between gaze and hand when taps are performed to indicate that they user is not able to interact with the tablet properly and propose a self-adapting response to bring the targets at better locations. When such a personalised model is not possible, our results serve as a baseline reference. In such scenario (for example in a public space when interaction is usually short and not regular), the adaptive system can detect if the users needs longer time to interact with it because of her navigation pattern (for example non native user will take longer to tap) and adapt its temporal threshold accordingly.

In terms of design, our findings highlight the fact users mainly focus on the top-left part of the screen: important information should therefore be located towards this direction not to be missed out by the users.

4.8 Conclusion

This chapter focused on the correlation between gaze and taps on tablets, tap being the main event happening during the manual interaction with the tablets, and a counterpart to the mouse clicks. We selected three Internet activities as a context for the data collection on gaze and tap inputs. Going on Internet is a natural and common activity on tablets, and there is a plethora of tasks and study method examples in research literature (cf. Chapter 2.3). These tasks were: a search task (containing search questions derived from cases found in literature), a shopping task, and a link-following game task (that was essentially designed to generate more tap data). By evaluating the relationship between the distance and the starting moment of the fixations (relative to the tap), we confirmed an expected result observed from gaze-mouse studies [57]): when tapping a target, gaze acquired the target before the tap was performed. On average we found that the shortest distance between gaze and taps was about 159 pixels, and it happened 338 ms before the tap occurs. This result serves as a baseline representing the typical healthy users behaviour. Subjects presenting a physical or cognitive disability may therefore be detected against this baseline and allow the interactive system to adapt itself to meet their requirement.

We studied a specific fixation, $F_{Closest}$, defined as the fixation happening *before* the tap occurs, at the closest distance to the tap. $F_{Closest}$'s starting moment to tap and distance to tap were consistent with the results found while studying the relationship between distance and time (median start moment to tap: -316 ms, median distance to tap: 51.5 pixels). $F_{Closest}$'s spatial distribution in the two dimensions of the screen revealed two areas of interest: around the tap location (for most of fixations), and a small cluster on the upper-left corner of the screen. We interpreted this second area of interest as the consequence of a known behaviour from Internet users to focus mainly in the upper-left part of the webpages (F-pattern [144]). Nevertheless, the median value of $F_{Closest}$'s position was very close to the tap point (-8.7,-12.7) pixels. This result highlight the importance of the application content's layout. It suggests that users would preferably visualise the information situated on the top-left of the screen before making their validation (by tapping).

We acknowledged disparities between participants, tasks and target types. Individual style of interaction had an impact on the correlation between gaze and tap, as users deployed different strategies to tap. For example, some users originally prepared a tap

but meanwhile, they kept searching the page with their hand ready to tap at one location, and eventually performed the tap without having to visually acquire the target again with precision. During typing, some participants were more likely to keep their eyes to the keyboard and monitor their actual tap, while others favoured constantly going back and forth between the keyboard and the input field to monitor both their tap on the keyboard and the effect on the input field during a typing sequence. Some participants did not look at the keyboard at all while typing. This behaviour is clearly shown by the impact of the typing skill level on the euclidean distance and vertical distance between $F_{Closest}$ and taps: the distances increased when the participants demonstrated a better typing skill level, since they did not require to monitor the keyboard to type and focused on the screen's location the typing string were on instead. We found that the spatial difference between gaze and tap was more important than the temporal difference. The nature of the task also influenced the correlation between gaze and hand: effect learning and anticipation inherent to a specific activity impaired the correlation. When users are more familiar with the application, less tapping errors occur (less distance between gaze and tap).

5

Correlation between Gaze and Stationary Hand Event

5.1 Introduction

Interaction with tablets relies mainly on the touch modality. Therefore, in a classic scenario of tablet interaction, it is possible to perceive the user's choices only once the decision has been realised via a tap. However sensing their hand movements in the depth space above the tablet not only brings new interactions methods [35, 37, 84, 148, 203] but can also inform much more on the user's cognitive process and mental state such as frustration [3]. Exploration of the tap and gaze correlation showed that gaze precedes touch [210] (cf. Chapter 4). However, to our knowledge, other parts of the hand events involved in the tapping process are not yet studied. We focus on the gaze behaviour during the occurrence of specific hand events for which the hand marks pauses. These stationary hand events (*hover* or *dwell* depending on the context, cf. Section 5.3 for how we categorise them) may happen before a tap is performed, the preparation of the taps. Both gaze and hand accompany the human cognitive process, in memory retrieval in particular [109, 188]. So understanding how gaze and hand behave before a tap can provide the machine indications on the user's cognitive process and give them matter to anticipate the adequate following steps in line with the user's needs.

In our work, we treat the stationary hand events as one possible indicator of hesitation (in the sense of decision making common cognitive process, rather than as a pathological

state). We focused on this example of user's cognitive behaviour because we wanted to propose an interaction method based on indecision in which second choices can be proposed to the user when the first choice has been discarded. We are interested to know where, during these events, the hand is located according to the display location the user is gazing at. We also investigate how this relationship between gaze and hand can inform on the indecision the user experiences while selecting targets. In this chapter, we consider "indecision" as the cognitive state of not being able to make a clear choice, without serious impact to the user or her actions (whereas "indecisiveness", described in [154], indicates a state where the user experiences "*decision delay, worry and regret*").

In this chapter, we present a data collection of gaze and hand positions while playing a "Memory Game". The choice of this game, as explained later, has been particularly made to study how gaze behaves while users keep their hand steady above the tablet. Besides, it offers a way to generate some cognitive activity from the users, and a definite framework for our analysis. The hand movement pattern is similar to the gaze movement: eye fixations can be assimilated to the stationary hand events, and saccades (quick eye movements between fixations) as when the hand is moving. Therefore, we propose to rely on this analogy to detect the stationary hand events, inspired by the algorithms commonly used in the field of eye tracking to detect fixations: dispersion and velocity algorithms [170]. Afterwards, we describe the spatial relationship between gaze and hand during stationary hand events for the different parts of the screen, and explain how this relationship changes when the user faces indecision.

5.2 System Design

We designed a system to collect and analyse data to understand the gaze and hand correlation during stationary hand events, part of the target selection process on a touch device. We paid attention to propose an application that generated enough matter for the participants to require taking decision, and forbore limiting the participants' engagement with an abstract study.

5.2.1 Content

Our data collection includes: 1) the eyeballs position and gaze samples provided by the eye tracker, 2) the hands position provided by Leap Motion, 3) the tap samples from

our tracking application based on the Microsoft Raw Input API and 4) the game event information logs (i.e. when a tile has been flipped, when a pair has been matched).

5.2.2 Context

For the context of our data collection, we implemented a "Memory Game": 12 shuffled pairs of pictures, shown face down, the player had to match by flipping them (tapping). This choice was driven by our interest in understanding the users' decision making process, while maintaining a joyful and motivating user experience. This game met these two criteria, and solely solicited the participants' memory (preventing difficulties brought by other factors such as language, general knowledge, etc.). Besides, it was supported by a very simple interface. Having the same interface across participants, as well as limiting the scope of actions (tap to flip a tile, match a pair) helped with framing a clear reference for further data exploration. The tiles (304 × 304 pixels) were arranged in 6 columns by 4 rows. When a pair was found, it did not flip back and remained in the game.

We did not constrain the participants with using only their dominant hand. They were free to interact the way they wanted to keep the naturalness of the experience. The resulting hand data is convereted as a single "flow" by selecting the closest hand to the tablet, as explained in Section 5.3.1.2.

5.2.3 Apparatus

The game was played on a Microsoft Surface Pro 4 (screen dimensions 260.28 × 173.52 mm, 1824 × 1216 pixels). The eyeballs and gaze positions were collected using a Tobii EyeX sampling at 60 Hz. The hands position was collected with a Leap Motion running at approximately 110 Hz. We designed a 3D-printed support to hold the tablet and both sensors in place. The digital modelling of the support has been conceived with Rhino[1], and we printed it with a Formlabs Form 2 stereolithography (SLA) printer[2]. It laid on a table (90 cm height) and has been devised so that each sensor could track without interfering with each other (infra-red emissions) and so that their respective fields of view cover the targeted body parts during playing. The position of the elements of the apparatus on the support board is indicated in Appendix B, Figures B.1 and B.2. Prior the data collection, we verified the data quality of the sensors working together on the

[1] `https://www.rhino3d.com/` (last accessed Jan. 2020)
[2] `https://formlabs.com/3d-printers/form-2/` (last accessed Jan. 2020)

support in a short feasibility study including 7 participants (volunteers from the same office or close friends - age 26.6±7.6, 3 females) who were asked to perform only one game round, using their dominant hand index. We asserted the data coherence in this feasibility study by observing a replay of the estimated gaze points and index positions in a representation of the tablet display (Figure 5.1 illustrates the final result of the display for one of the participants).

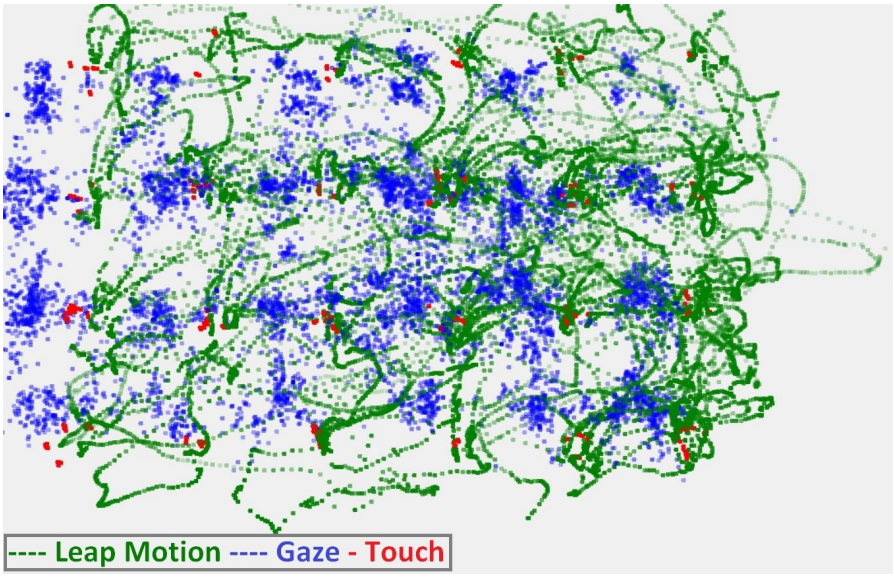


Figure 5.1: Data visualisation for the feasibility study for one participant.

Regarding the software component of the apparatus, we designed a C# application to manage the sensors and retrieve their logs, as well as to launch the different elements of the data collection (calibration programme, game, drift assessment). Each sensor's API provided timestamps that we synchronised with the system clock via the manager application.

Figure 5.2 illustrates the apparatus deployed in a public space. Participants typically stood up about 66 cm away from the tablet centre.

5.2.4 Protocol

Before playing the game, the participants filled a consent form, and we assessed their hand laterality and their dominant eye (triangle test[3]). The participants were then introduced to the game with a demonstration version (3 ×2 abstract figures in larger size pictures, cf. Appendix B, Figure B.3).

[3] `http://www.allaboutvision.com/resources/dominant-eye-test.htm` (last accessed Jan. 2020)

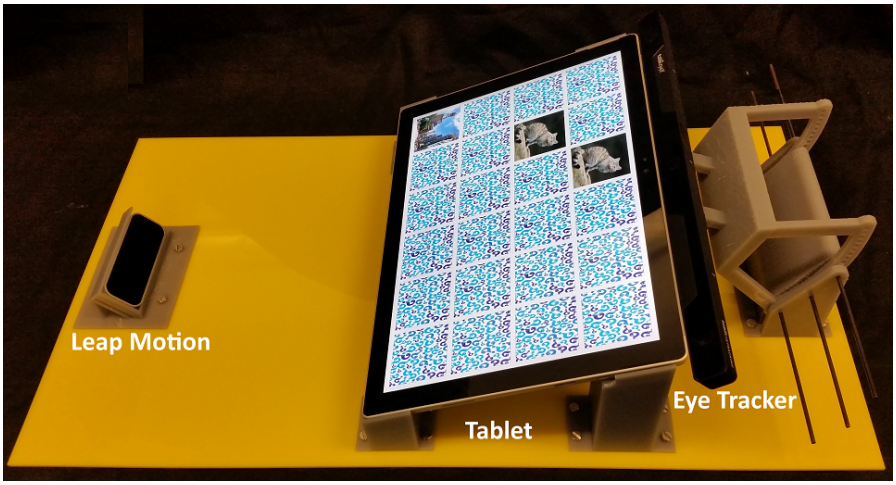


Figure 5.2: System used during the data collection of stationary hand events.

A 5-point eye tracker calibration was performed before the data collection (accuracy of 0.73°±1.9) as explained by Chapter 3.1. Three increasing difficulty levels were played. Level 1 showed pictures of various animals, objects or landscapes which disparity left few chances for mistaking one with another (at least semantically, cf. Appendix B, Figure B.4 to see their content and position on the game). Level 2 only included pictures of trees in different landscapes (cf. Appendix B, Figure B.5): a representation of the same conceptual object in different variations (colours, shape, environment of the picture). Level 3 only contained pictures of close-up sea surfaces (cf. Appendix B, Figure B.6): a conceptual object harder to differentiate under several representations, we tried to find textures and colours that remain similar enough to be difficult to distinguish at first sight, while still allowing the completion of the game. To avoid learning effects, each pair of pictures' locations were different from one level to another. The participants' hands movements were also video-recorded. To finish, a 5-point accuracy test was run (0.79°±4).

5.2.5 Participants

To recruit participants, we set up the study apparatus in a public area of the Lancaster University campus. Flyers (shown in Appendix B Figure B.7) were displayed at the vicinity of this public space, but we mainly directly approached people and asked them if they were willing to give some time to play a Memory Game for the study (which took about 15 minutes on average), and proposed snacks as a token of gratitude. As for the previous data collection (Chapter 4.3.4), we did not discard participants unless the calibration with the eye tracker was not possible at all (details in Chapter 3.1.4).

In total, 117 participants played the game (49 female, age 26±8.6). Most of them were right-handed (103) and their right eye was dominant (83). Since English comprehension was not important for the study's tasks, we did not record information regarding how English was mastered by the participants, even if for a large part of them, English was not their native language. After discarding the trials resulting in a poor data collection (either from the hand tracker or the eye tracker) we kept 177 trials across 71 participants.

5.2.6 Eye Movements Classification

In a post hoc step, we extracted the fixations from the gaze data by running a dispersion algorithm (IDT algorithm [170]). The temporal threshold used in the algorithm was 100 ms for all participants. However, the spatial threshold we set for the dispersion detection varied among them. We computed the equivalent length on the screen of 2° of visual angle, based on the average distance, collected during the game, between the tablet and the participant.

5.2.7 Hand Events Classification

We focused on the stationary events of the hands that reflect the potential choices the participants considered. To retrieve these events, we adapted algorithms from eye tracking based on the similarity between the patterns of the gaze and hand movements. Section 5.3 is dedicated to the details of these algorithms and their evaluation.

Table 5.1: Stationary hand events (hovers and dwells) number and round duration percentage for the validation subset.

Validation subset element index	Number of hovers / dwells	% of round duration
1	42 / 3	12
2	64 / 36	33
3	41 / 16	25
4	76 / 30	28
5	41 / 5	12
6	33 / 36	30
7	66 / 13	27
8	56 / 23	48
9	29 / 69	34
10	46 / 14	36

5.3 Stationary Hand Events Detection

5.3.1 Data Preparation

From our valid data collection, we extracted a subset of 10 game rounds randomly chosen for validating the algorithms. In a second step, another subset of 10 game rounds has been picked up for testing the algorithms.

5.3.1.1 Ground Truth

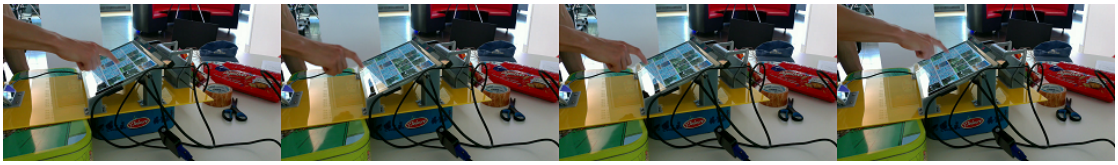
The only medium from the data collection we could access for preparing the ground truth files were the videos of the hands. We annotated the videos of the training subset by observing when the hand stayed in a stationary position. Based on this annotation, the ground truth files were generated to match the **events** of the videos and to indicate when the hand was in a stationary position. To align data correctly, we defined a reference duration (the starting time was the start of the Leap Motion for the concerned round, the ending time was the last tap performed at the concerned round).

From the observation made during the data collection and the videos, we could clearly observe different behaviours among participants. On of the most striking personal features were related to how the hands are kept above the tablet and the change of hands during the interaction. Figure 5.3 illustrates the first difference: some participants preferred keeping their hand away from the tablet as much as possible, hence moving above the tablet's display only when a tap was required (Figure 5.3a), whereas others kept their hand above the surface, either in a relatively still position to move only when tapping or "following" the gaze while skimming the tablet's content (Figure 5.3b. This trends were not systematical: some participants were showing both type of trends during the interaction. A visual classification based on the video we took during the data collection indicates that on average during the interaction with the tablet in this context of playing Memory Game, most participants maintained they hand over the display (69 % of them, 81 participants) while more than a quarter of them (29 %, 34 participants) did not have a specific pattern and sometimes kept the hand above the display and sometimes retracted it back closer to their body. Only 2 participants mainly retracted close to their body during most of the interaction.

The other personal feature we observed concerned the switch of hands during interaction. Despite the fact participants reported a dominant hand before the interaction started, we

noticed that all participants did not necessarily interact with the tablet solely with that dominant hand: some participants were more likely to switch hands during interaction. We associate this behaviour with two factors: fatigue and/or ease of reach (less effort is required to reach the left side of the tablet with the left hand and the right side of the tablet with the right hand). Video analysis showed than on average during interaction, participants solely use their dominant hand (90 %, 105 participants), and few users preferred using both hands: 6 participants (about 5 %) used both hands to reach each side of the tablet, while 6 others used one hand a time alternatively.

(a) The hand is returning to a base position near the body between taps.



(b) The hand is staying near the tablet screen and can follow the gaze path.

Figure 5.3: Screenshots of different hand moving trend sequences during interaction.

When observing the personal feature on a game level basis, we noticed that often, participants changed their interaction pattern during playing. For example, some participants started playing the first level with keeping a close distance between their hand and the tablet, then started to retract the hand close to their body for the last level - and some participants did the exact opposite. The conclusions from this change of behaviour are difficult to draw, because it is not clear which factors influences this change. Fatigue and/or mental load are two candidates but there is no evaluation possible to assess which of them are the most influential on the change of behaviour. From our population sample, 26 % of the participants (30) changed their personal pattern (as keeping the hand above the surface of the tablet or retracting it) and 13 % of them (15) changed hand pattern for the left/right hand usage during the whole interaction (for instance starting with both hands then only one hand).

5.3.1.2 Hovers and Dwell Definitions

The nature of such data expectingly results in a highly skewed classification, as suggested by Table 5.1: during interaction the hand was most of the time in movement rather than stationary (from our validation subset, stationary hand events lasted between 12 % and 48 % of the overall duration of the round).

In order to detect the type of the stationary hand event (hover or dwell), we evaluate *where* the hand *was targeting* at, based on the hand position retrieved from the Leap Motion, and on the dominant eye position retrieved from the eye tracker. During the duration of an event, we compute the *projection* of the hand from the eye, as illustrated by Figure 5.4. From the Leap Motion, we select the finger tip point that was the closest to the tablet's top (lowest Y-value in the tablet coordinates) in order to compute the projection. If the projection is on the tablet, the stationary hand event is considered as *hover*, otherwise it is labelled as *dwell*.

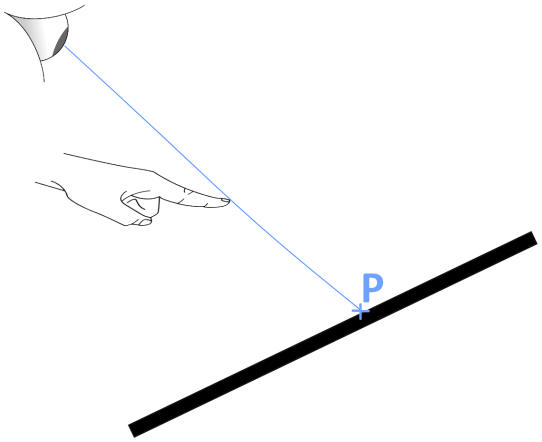


Figure 5.4: Projection (P) of the hand on the tablet from the user's eye perspective.

5.3.2 Algorithms Presentation

Since we retrieved the palm's and fingers' tip positions with the Leap Motion, we first need to filter the data to work with a single point representing the hand motion in space and time. We select the palm's position, as it represents well the hand motion (if the hand was still the palm was still). Besides, we noticed while observing the participants interacting with the tablet, that although their palm remained relatively stationary, they sometimes still moved their fingers meanwhile. We also realised that they did not systematically only use the same finger to point at the table during movement. Since

both hands can be used during interaction we always retrieve the palm position of the *closest* hand to the tablet surface.

Inspired by the algorithms proposed for eye tracking to retrieve gaze fixations (when gaze stays in close-to-stationary position), we evaluate three algorithms for extracting stationary hand events that we ran on the palm position data across all levels of the game (the hover detection is an event detection and therefore do not depend on the level).

- Identification by Dispersion Threshold algorithm (**IDT**): the dispersion is defined as $d = [max(x) - min(x)] + [max(y) - min(y)]$ as mentioned in [94]. The algorithm requires two thresholds: a temporal threshold corresponding to the minimum duration of what could be considered a stationary hand event, and a spatial threshold corresponding to the maximum value the dispersion can take to be considered a stationary hand event. For reference values, we select a spatial threshold of 5 mm, value coherent with studies on natural hand tremor [216] (maximum displacement of a finger tremor), and a temporal threshold of 100 ms, like gaze fixation detection explained in Section 5.2.6.

- Identification by Dispersion Threshold - Euclidean algorithm based on the Euclidean distance (**IDTE**): this algorithm is similar to IDT, but the dispersion was computed as the average Euclidean distance between the samples.

- Identification by Velocity Threshold algorithm (**IVT**): this algorithm computes the instant velocity between the hand motion samples. It also requires a spatial and a temporal threshold. The temporal threshold has the same role as for IDT and IDTE, and the spatial threshold corresponds to the maximum speed the hand movement can reach to be still considered a stationary hand event. The reference value for the spatial threshold is 80 mm/s, similar to the suggestion formulated by Vogel and Balakrishnan [202] as an indicator of pause velocity in their study of distant freehand pointing.

In the remaining part of this chapter, we will write the abbreviation "Tt" when referring to the temporal threshold, and "St" when referring to the spatial threshold.

5.3.3 Algorithms Performance

For each algorithm, we generate a file on the same format than the ground truth file: each line corresponds to an event, and indicates a hover or a dwell, depending on the finger's projection from the eye. The technique to extract the starting and ending time reference, as well as the hand's projection onto the tablet's area, are the same than those introduced in Sections 5.3.1.1 and 5.3.1.2. We generate a file for different values of the algorithm's thresholds.

When comparing the ground truth files with the algorithms files, we classify the events as *correct*, when an event from the algorithm matches at least one event of the algorithm (timely), *deleted* when it is not the case, and *inserted* when an event from the algorithm does not match an event from the ground truth. This terminology is taken from the work on activity recognition presented by Ward et al. [207], the respective equivalent in binary classification statistics are *true positive*, *false negative* and *false positive*.

As a consequence of the expected classification skewness mentioned in Section 5.3.1.2, *accuracy* is an inadequate metric since it is relevant when a binary classification is expected to be fairly balanced.

The first binary classification metric we report is the *recall* computed by equation (5.1). The recall rate indicates how good the classifier performs to find the true positives. However, by itself, it is not complete as the algorithm can produce a high recall rate but also find a lot of false positives. Therefore, the choice of a good algorithm depends on the scenario for which stationary hand events need to be detected. In other words, since a perfect classifier is unrealistic, two concrete cases need to be considered: (1) "minimising false negatives" classifier: it is more important for the application that **all** stationary hand events are retrieved at the cost of getting some errors of the classifier detecting stationary hand events when there is none, or (2) "minimising false positives" classifier: the application need to retrieve events that are **only** stationary hand events at the cost of not detecting some. The first case aforementioned is represented by the recall rate.

$$recall = \frac{Correct}{Correct + Deleted} \tag{5.1}$$

The metric we can report to interpret the second case aforementioned is the *precision* defined by equation (5.2), a good indicator of how an algorithm strictly identifies actual events since it is measuring the ratio of true positives from the classifier's retrieved

instances.

$$precision = \frac{Correct}{Correct + Inserted} \tag{5.2}$$

Recall and precision are two metrics than should be reported together to understand how the classifier behaves. An ideal classifier would give a high recall rate with a high precision rate (both close to 1). Since this is hardly the case in a real-case scenario, the choice of a classifier will often be the one that offers the best compromise between the two metrics.

A common metric to evaluate this comprise is the F1 score [196], defined as the harmonic mean of the precision and the recall rates, formulated in equation (5.3), that we report as well in our results.

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{5.3}$$

We present, in the following, the metrics results for the validation stage of our algorithms' evaluation - for both types of stationary hand events, then for dwells only and lastly for hovers only.

5.3.3.1 Dwells and Hovers

Figures 5.5 and 5.6 show how the precision and recall rates evolve together for the IDT algorithm depending on different values of Tt and St respectively. Plots for the other algorithms are given in Appendix C, Figures C.1 and C.2 for the IDTE algorithm and Figures C.3 and C.4 for the IVT algorithm.

For each algorithm, precision and recall are clearly opposite, when reaching a good precision, the recall is extremely poor, and vice versa. Moreover, with IDT and IDTE, the distribution between precision and recall when the thresholds vary is clear: the precision increases with Tt, and when St increases, better recall is observed. There is no obvious case of a threshold combination that would optimise both precision and recall altogether (which should appear towards in top-right corner of the graphs mentioned above).

Figure 5.7 shows the F1 scores for IDT found at different thresholds combinations. Illustrations for the other algorithms are provided in Appendix C, Figure C.5 (IDTE) and C.6 (IVT). All best scores per algorithm are reported in Table 5.2. Among the three algorithms, the highest F1 score (0.818) is obtained with the IDT algorithm, for Tt 165 ms and St 12.8 mm.

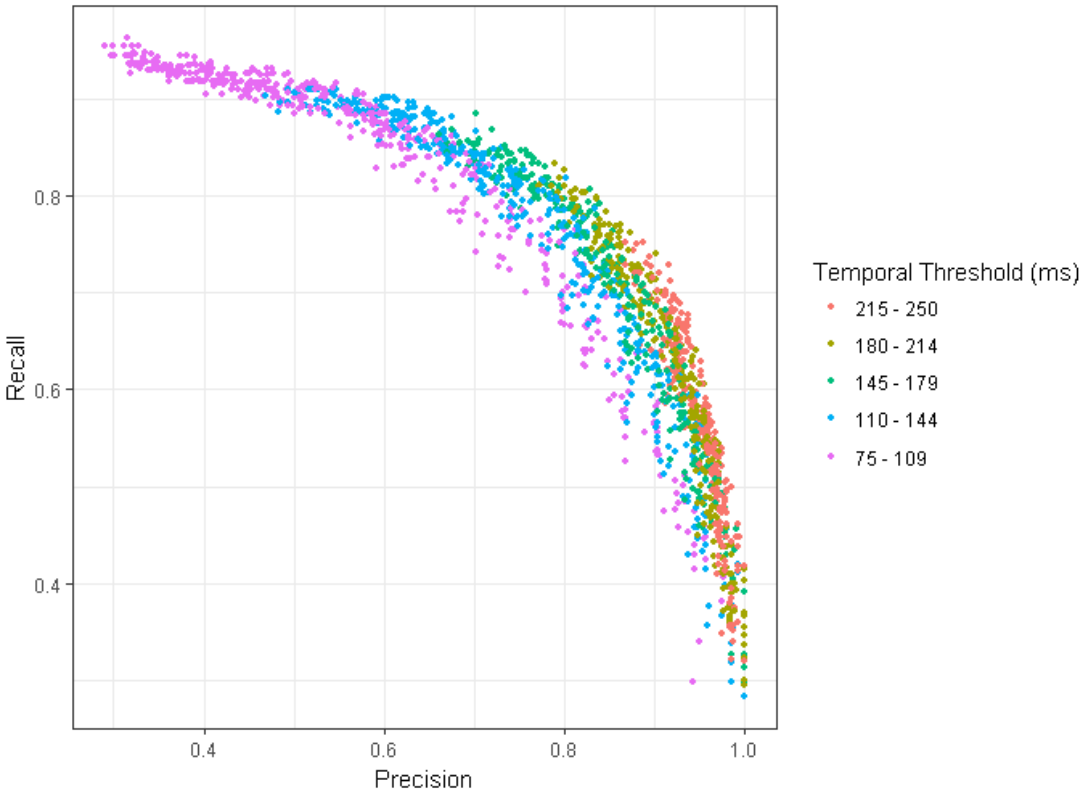


Figure 5.5: IDT Precision-Recall space for the IDT algorithm (grouping by Tt).

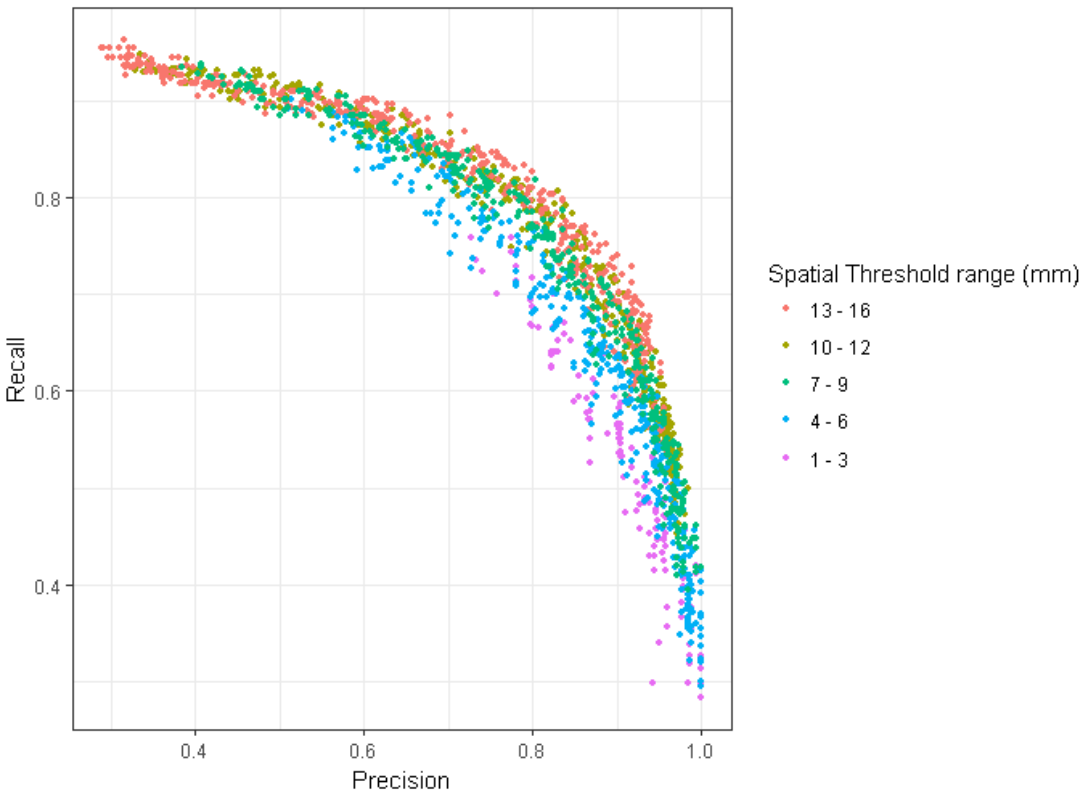


Figure 5.6: Precision-Recall space for the IDT algorithm (grouping by St).

Table 5.2: Best F1 scores per algorithms (dwell and hover).

Algorithm	F1	Precision	Recall	Tt	St
IDT	0.818	0.829	0.807	165 ms	12.8 mm
IDTE	0.813	0.897	0.744	200 ms	9.2 mm
IVT	0.811	0.813	0.809	120 ms	76 mm/s

5.3.3.2 Dwells Only

We report the previously introduced metrics for the case of dwells only (stationary position of the hand outside the tablet's volume). The distribution between precision and recall for IDTE is illustrated by Figures 5.8 and 5.9 against the variation of Tt and St respectively. Appendix C contains the plots for IDT (Figures C.7 and C.8) and IVT (C.9 and C.10).

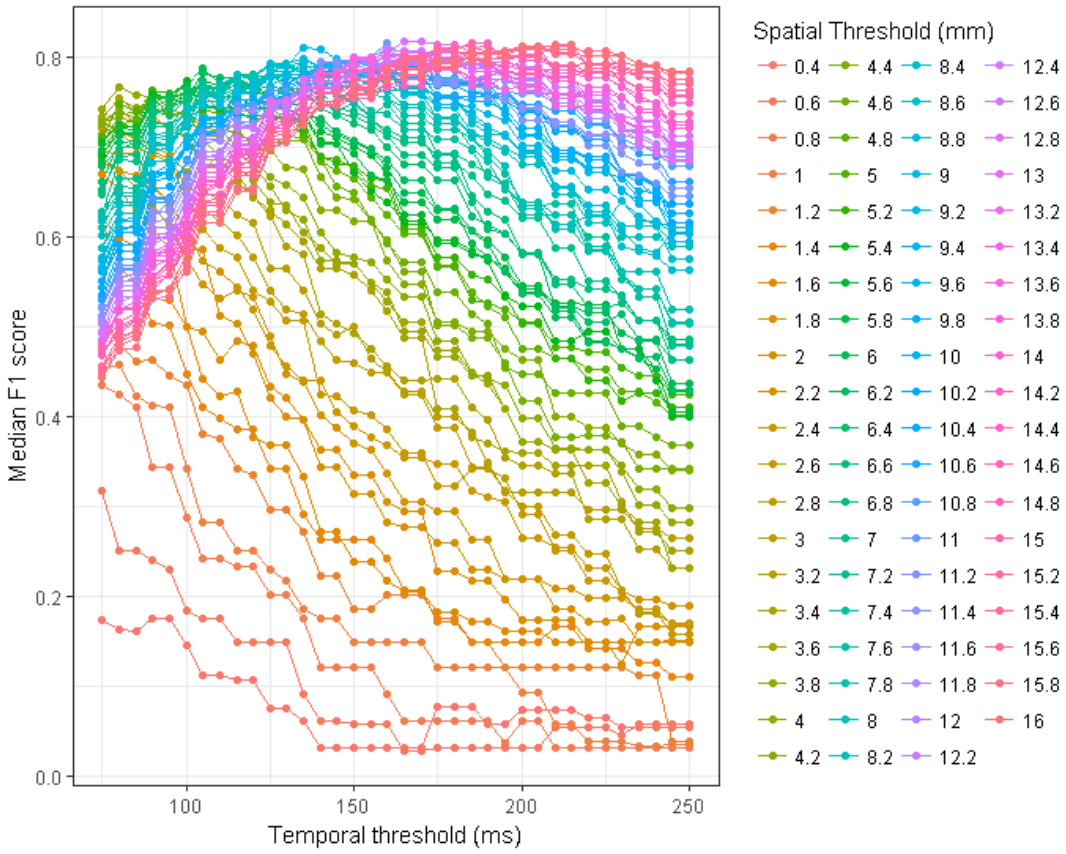


Figure 5.7: F1 score for the different combinations of thresholds of the IDT algorithm.

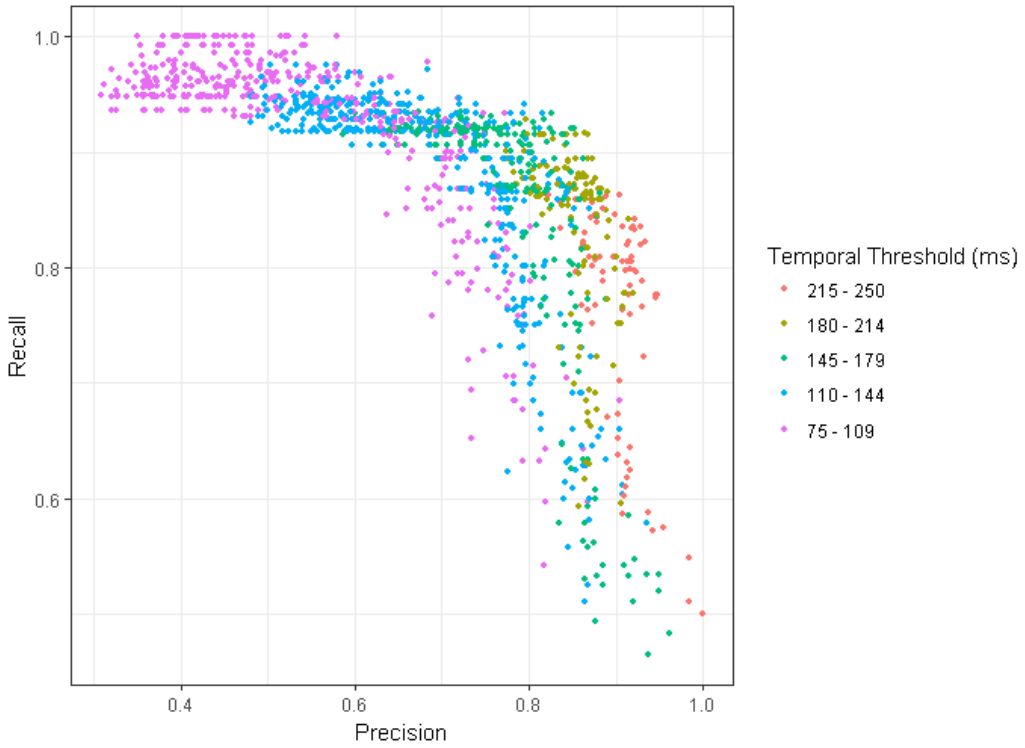


Figure 5.8: Precision-Recall space for the IDTE algorithm (grouping by Tt, dwells only).

This distribution evolves according to Tt and St in the same way described previously for all types of stationary hand events. However, the case of dwells only shows a slight difference: the recall rate keeps on a relatively high value even when the precision is at its maximum (about 0.5).

Figure 5.10 shows the F1 scores computed with different thresholds combinations when dealing only with dwell events for the IDTE algorithm. Cases for the IDT algorithm (Figure C.11) and for the IVT algorithm (Figure C.12) are illustrated in Appendix C. The best F1 scores per algorithm are reported in Table 5.3. The overall best F1 score is 0.910 for IDTE (Tt = 160 ms, St = 9 mm).

Table 5.3: Best F1 scores per algorithms (dwell only).

Algorithm	F1	Precision	Recall	Tt	St
IDT	0.897	0.943	0.855	185 ms	12.8 mm
IDTE	0.910	0.893	0.929	160 ms	9 mm
IVT	0.847	0.885	0.812	105 ms	66 mm/s

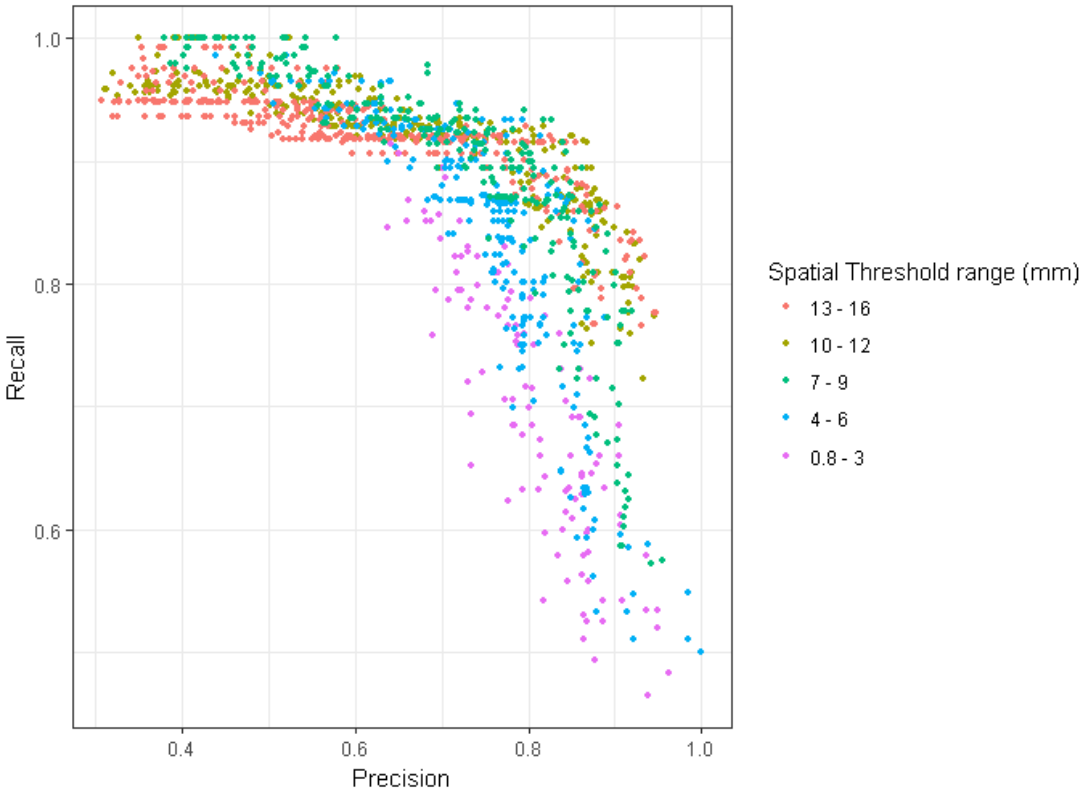


Figure 5.9: Precision-Recall space for the IDTE algorithm (grouping by St, dwells only).

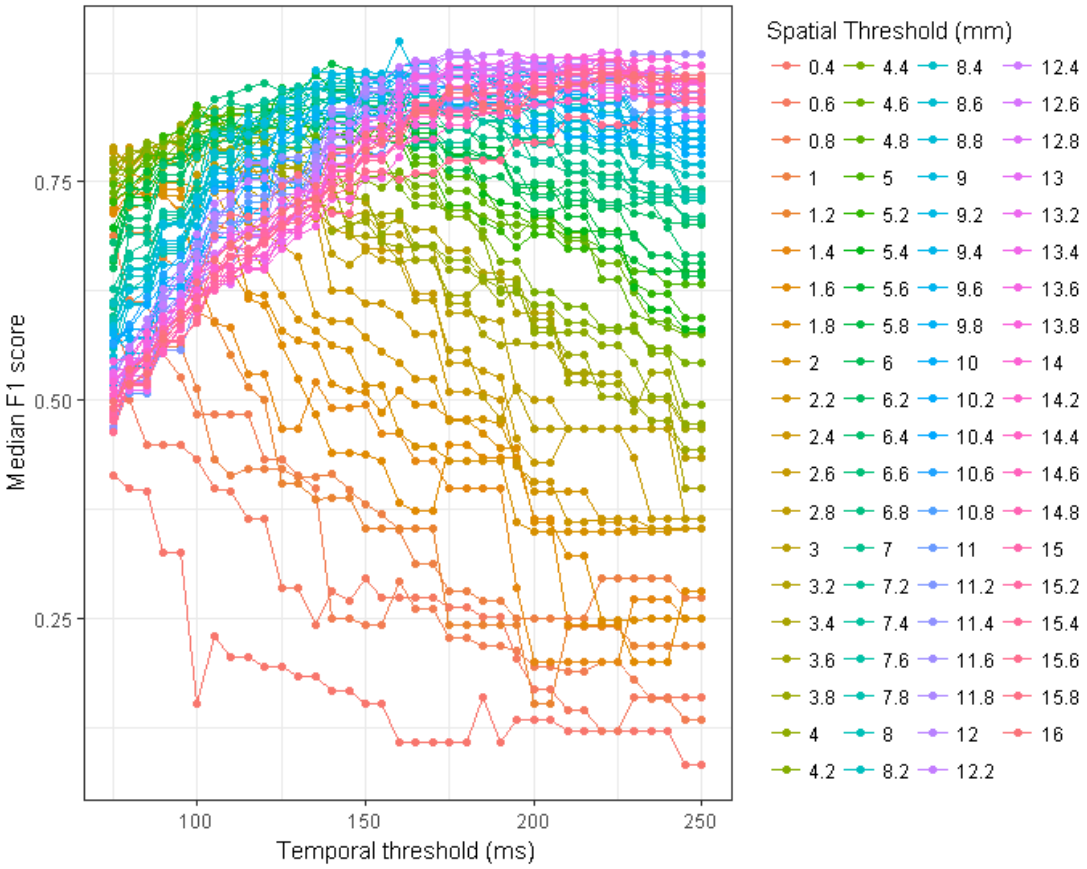


Figure 5.10: F1 score for the different combinations of thresholds of the IDTE algorithm (dwells only).

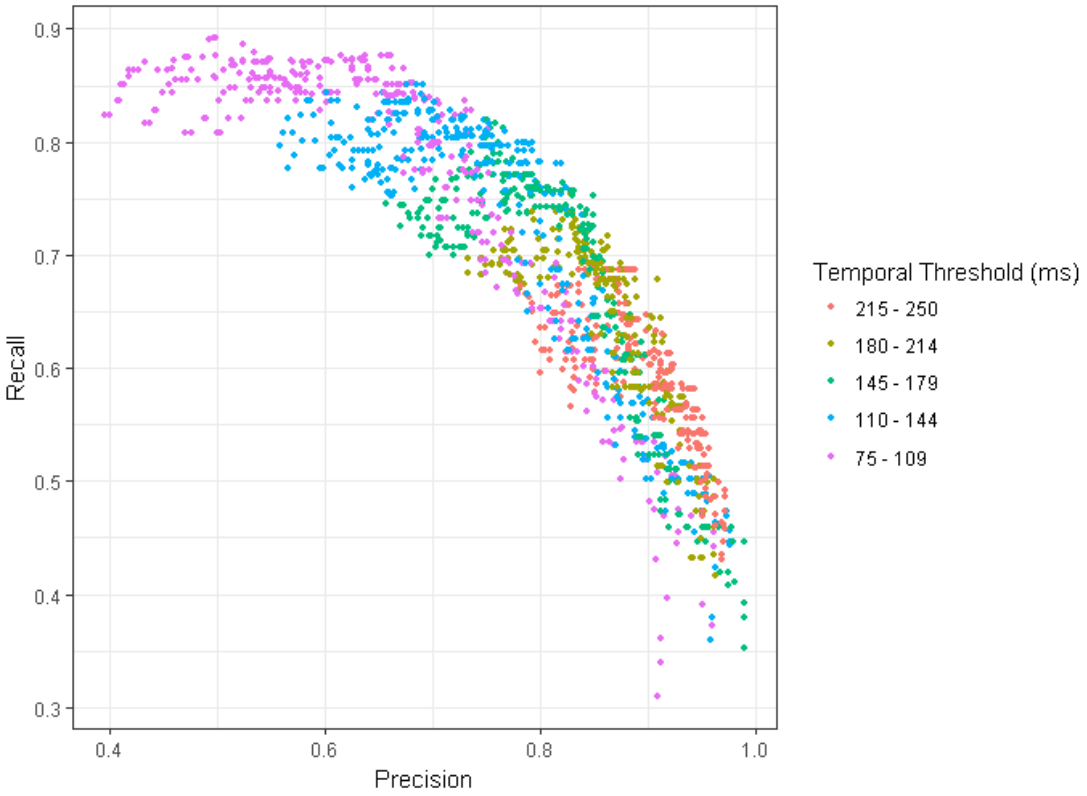


Figure 5.11: Precision-Recall space for the IVT algorithm (grouping by Tt, hovers only).

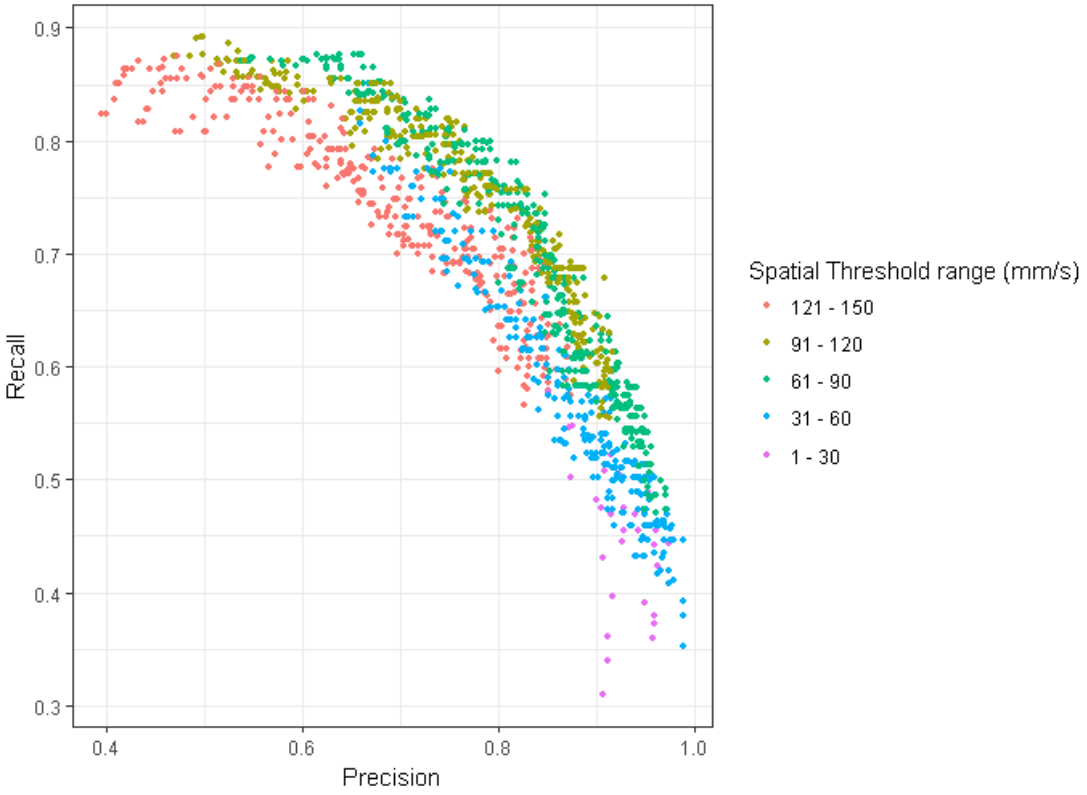


Figure 5.12: Precision-Recall space for the IVT algorithm (grouping by St, hovers only).

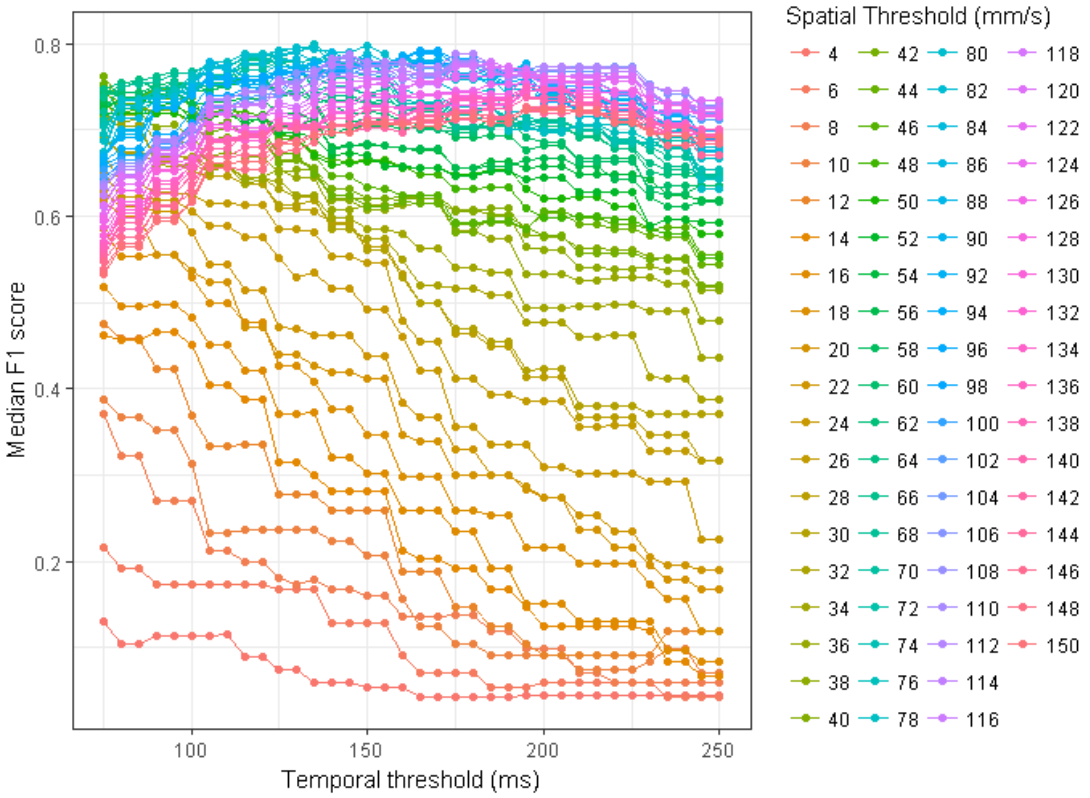


Figure 5.13: F1 score for the different combinations of thresholds of the IVT algorithm (hovers only).

5.3.3.3 Hovers Only

From the event counts given in Table 5.1 it is clear that the results from hovers and dwells together are mainly influenced by hovers since they are, except for two subset elements, the main type of events for the stationary hand events.

The precision/recall distributions for the IVT algorithm on hovers only, per Tt and St respectively, are plotted in Figures 5.11 and 5.12. The equivalent plots for IDT (Figures C.13 and C.14) and IDTE (Figures C.15 and C.16) are included in Appendix C. The impact of Tt and St is the same as before.

The F1 score of hovers for the IVT algorithm with different values of Tt and St is given by Figure 5.13, and in Appendix C by Figures C.17 (IDT) and C.18 (IDTE). The best F1 score, 0.819 (Tt = 220 ms and St = 10.4 mm) is found for the IDTE algorithm, among the values for each algorithm reported in Table 5.4.

Table 5.4: Best F1 scores per algorithms (hover only).

Algorithm	F1	Precision	Recall	Tt	St
IDT	0.809	0.866	0.758	210 ms	16 mm
IDTE	0.819	0.879	0.766	220 ms	10.4 mm
IVT	0.799	0.818	0.781	135 ms	76 mm/s

5.3.3.4 Test

We test our algorithms on a random subset of the data collection, containing elements that have not taken part of the subset from which we validated the algorithms. Table 5.5 summarises the metrics' values of the best cases we found in the validating set, described in the previous sections.

Table 5.5: Testing set results.

Case	F1	Precision	Recall
Dwells + Hovers	0.705	0.652	0.767
Dwells only	0.733	0.669	0.810
Hovers only	0.731	0.690	0.777

The results we get for the testing set are acceptable, but not as good as the validating set. The explanation we can give for this observation is that the algorithms are sensitive to the personalised way of interacting with a tablet. To compare the results from this testing set with the results obtained with the validating set, we expose the best F1 score values of the testing set and summarise them in Table 5.6.

Table 5.6: Testing set results for best F1 values.

Case*	Algorithm	F1 (Precision, Recall)	Tt	St
D+H	IDT	0.767 (0.814, 0.726)	185 ms	11.2 mm
D+H	IDTE	0.766 (0.867, 0.686)	200 ms	7.2 mm
D+H	IVT	0.718 (0.841, 0.626)	175 ms	68 mm/s
D	IDT	0.778 (0.814, 0.744)	195 ms	10.4 mm
D	IDTE	0.777 (0.805, 0.750)	200 ms	7.2 mm
D	IVT	0.728 (0.746, 0.711)	105 ms	42 mm/s
H	IDT	0.808 (0.840, 0.778)	195 ms	11.2 mm
H	IDTE	0.819 (0.821, 0.816)	160 ms	6.2 mm
H	IVT	0.786 (0.840, 0.739)	140 ms	50 mm/s

(*)D = Dwells H = Hovers

5.4 Relationship between Gaze and Stationary Hand Events

Our dataset contains 13443 stationary hand event samples and 46812 fixation samples. We construct our study based on the assumption that the stationary hand events happen **before** taps. They may not lead to the tap straight forward: for instance, if a user hesitates to tap, the hand hangs, then moves and probably hangs again before effectively tapping. As mentioned in Section 5.3, we also classify the stationary hand events in

two categories depending on whether they occurred strictly above the tablet's surface or not (hovers and dwells respectively). The stationary hand events are retrieved with the best performing algorithm for both dwells and hovers: IDT with a temporal threshold of 165 ms and a spatial threshold of 12.8 mm (cf. Table 5.2). Therefore, the resulting classification that combines the two aforementioned notions is summarised in Table 5.7.

The classification is based on the temporal order between the stationary hand events

Table 5.7: Classification of the hovers.

Leading to tap {L}	Hover {LH}	(5201, 38.7 %)
	Dwell {LD}	(2447, 18.2 %)
Not leading to tap {NL}	Hovered {NLH}	(3606, 26.8 %)
	Dwell {NLD}	(2189, 16.3 %)

and the taps (**L** stands for leading to tap, **NL** stands for not leading to tap) and the projection of the pointing finger onto the tablet display (**D** stands for dwell, **H** stands for hover, notions detailed in Section 5.3.1.2).

In the following parts of this section, when mentioning the stationary hand event position, it is assumed, unless not stated otherwise, that we actually mean the position of the *hand's projection* as described in Section 5.3.1.2.

When hovering or dwelling above the tablet, we intuitively do not expect the participants to manually point at the very same position with gaze to avoid occlusion. Therefore, we want to understand where participants keep their hand when it is marking a stationary event during the data collection.

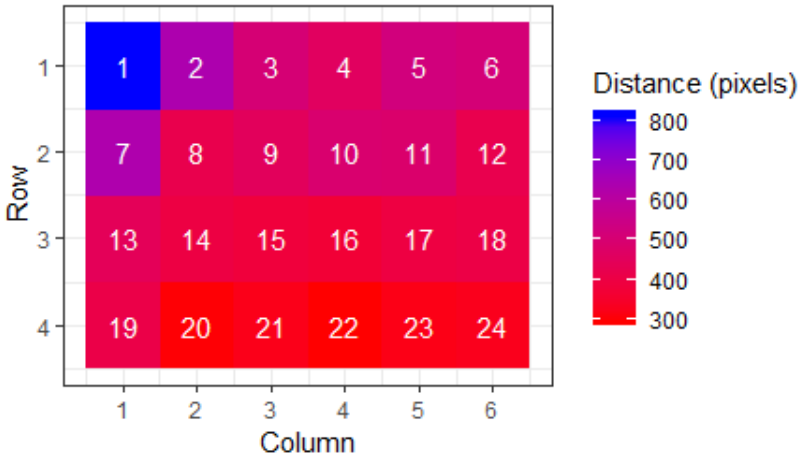


Figure 5.14: Relative median position between gaze and stationary hand events per tile.

We report the median average position of the stationary hand event relative to the gaze. Since our data collection context already divide the tablet's screen in several areas limited

by the game tiles, we analyse the aforementioned variable for these areas the participants looked at while playing the game. In a first step, we only keep stationary hand events during which gaze stayed on a same tile (78.4 % of all hovers). Figure 5.14 illustrates the median distance value between gaze and stationary hand events for each tile. It shows that the distance increases radially from the bottom centre-right of the screen (298±304 pixels on tile 22) to the top corners (813±310 pixels on the left and 512±467 pixels on the right). We explain this radial distribution from the participants' tendency to keep their hand at a minimal distance from their arm rest position, certainly to prevent arm fatigue. It also indicates that the participants used a "manual mapping" of the screen that was smaller than the actual projection of the screen at the stationary hand events depth level, and better aligned at the bottom centre-right of the screen where the hand, in its natural position, is the closest from.

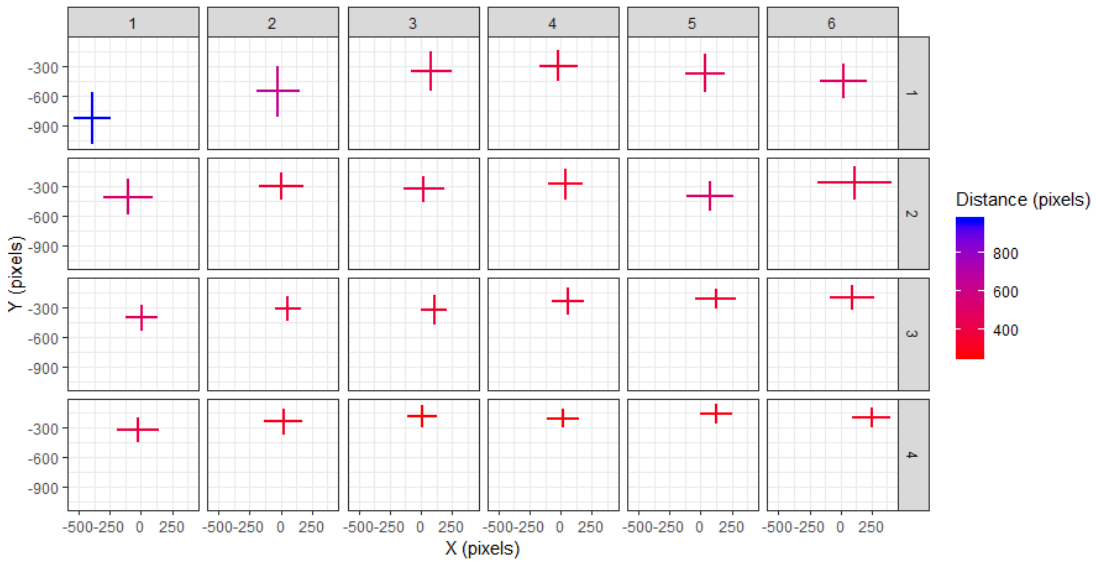


Figure 5.15: Relative median position between gaze and stationary hand events per tile for *left-handed* participants.

We investigated the role of the handedness in the distance between gaze and stationary hand events. As mentioned in Section 5.2.5, most people were right-handed. Figure 5.15 shows the horizontal and vertical distances between gaze and stationary hand events for each tile of the game for left-handed people, and Figure 5.16 for right-handed participants. We notice that, against the intuitive idea that there would be a clear difference between participants with different handedness, the distance distribution mentioned above does not change for left-handed participants: the biggest distance between gaze and stationary hand events is still found at the top-left corner of the screen, and there is still a radial distribution starting from the bottom centre. This behaviour may be explained by two

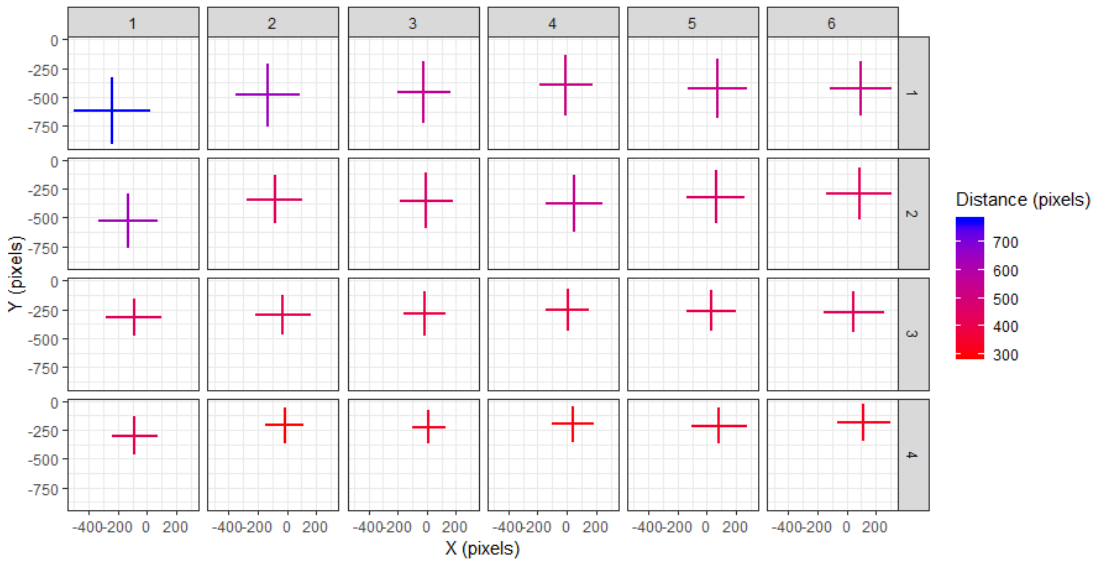


Figure 5.16: Relative median position between gaze and stationary hand events per tile for *right-handed* participants.

reasons: some participants mentioned they were left-handed but still interacted with the right hand during the games, and some participants (regardless their handedness), used both hands to interact (cf. Section 5.3.1.1). Therefore, in the following results, we do not distinguish the cases of handedness.

In a second step, we only focus on hovers, the stationary hand events inside the volume above the tablet's screen (**LH+NLH**), expecting them to be closer to gaze. As shown by Figure 5.17, the distance radial distribution over the screen we already introduced for all stationary hand events is still observed for hovers only. We notice that the median difference between gaze and hand positions, on the horizontal axis, increases at the edges and shifts approximately at the middle of the screen. On the vertical axis, this difference increases when the participants were looking towards the top border of the tablet, and is even bigger at the top corners of the screens. We interpret this as a trend for the participants to favour horizontal hand movements over vertical hand movements. The vertical position and distance between gaze and hovers (**LH+NLH**) in the one hand, and between gaze and dwells (**LD+NLD**) on the other hand are significantly different for each tile of the screen (Wilcoxon rank-sum test[4], $p < 0.01$ for every tile). As no scrolling on the tablet was required for interaction, participants probably tended to keep their hand close to their body (vertically) as a way to minimise limb efforts (explanation proposed by [4, 91]). These results are similar to observations made in several gesture preferences study works [68, 118] which indicate users favoured horizontal movements

[4]We did not assume normality of the data, therefore we used the nonparametric Wilcoxon rank-sum test to compare groups.

over vertical movements.

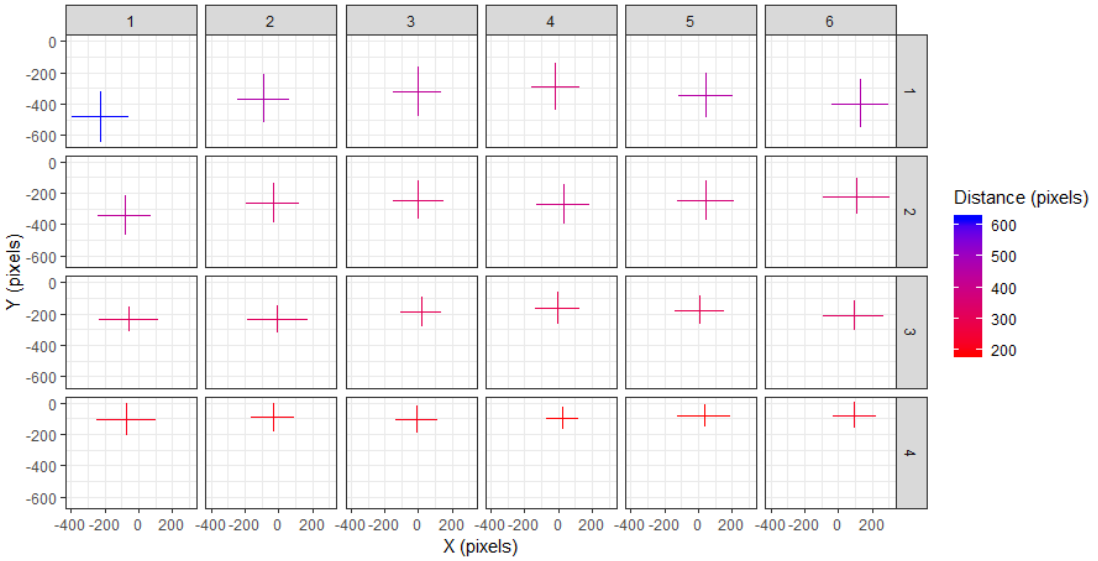


Figure 5.17: Relative median position between gaze and hover per tile.

We did not find a systematic pattern nor a significant difference between gaze and stationary hand events depending on if they were leading to taps (**L**) or not (**NL**).

5.5 Indecision and Gaze/Stationary Hand Events Relationship

Our primary motivation to adopt a Memory Game for the data collection was the willingness to work with tablet's user cognition evaluation. Accordingly, we wish to verify the hypothesis that the correlation between gaze and stationary hand events presents characteristics that reveal how participants were confident about their choices.

We evaluate indecision via the coarse approximation of pair matching failure on already seen elements. We only focus on **L** stationary hand events (because it indicated the participants were planning to tap), for tiles that had been seen before (to discard the exploratory phase of the game, when participants randomly flipped tiles to start the game) and that were the second element of the pair matching (to characterise the taps as "successful" or "unsuccessful").

We expected that when participants were facing indecision, the duration and/or the number of fixations during stationary hand events would be particular because they would indicate the participants were processing information in relation with their memory recall [131]. However, we do not observe a difference in the average number/duration of the fixations during stationary hand events leading to successful or unsuccessful taps (re-

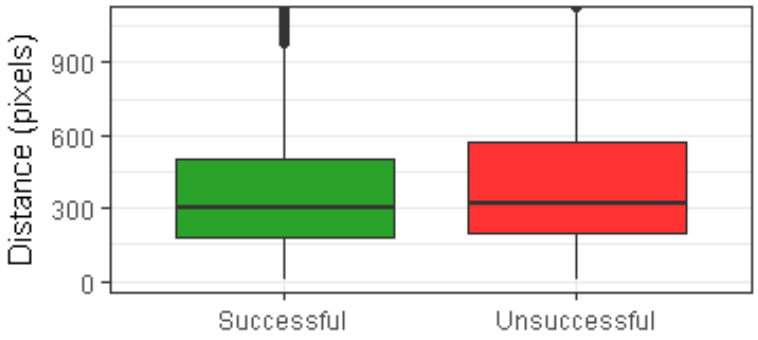


Figure 5.18: Median distance between gaze and stationary hand event positions.

Table 5.8: Median distance between gaze and stationary hand event positions (per game level).

Level	Distance between gaze and stationary hand events positions (pixels)	
	(successful)	(unsuccessful)
1	356±390	369±390
2	360±326	410±389
3	401±383	466±447

spectively 1.48±0.6 fixations for 256±181 ms and 1.48±0.66 fixations for 236±167 ms). Nevertheless, in the spatial domain, we observe a significant difference in the vertical position (Wilcoxon rank-sum test, W = 878860, p-value < 0.05) and distance (illustrated by Figure 5.18, Wilcoxon rank-sum test, W = 986870, p-value < 0.05) between the gaze and stationary hand events depending on the tap success or failure. The hand's distance and vertical position are closer to the gaze point (ΔY = -265±365 pixels, distance = 383±367 pixels) for stationary hand event that led to successful taps compared to unsuccessful taps (ΔY = -289±374 pixels, distance = 412±376 pixels).

Since our data collection context contained several levels of increasing difficulty, we suspect the observation made for the general case above to be dependant of the game level. Indeed, we predict indecision may be stronger when the difficulty of the task is higher as it seems a natural behaviour. Table 5.8 summarises the distance between gaze and hovers, for hovers leading to successful and unsuccessful taps, which is also illustrated by Figure 5.19. From these results, two observations can be made. Firstly, the distances between gaze and the stationary hand events are significantly different (for the events leading to successful and unsuccessful taps) only for the two *last* levels (Wilcoxon rank-sum test, W = 108780, p-value < 0.05 for level 2, W = 100620, p-value < 0.05 for level 3), which bore stronger cognitive activities. Secondly, the difference between the distances of each cases increased with the level (Δ= 13 pixels for level 1, Δ= 50 pixels for level 2

and Δ= 65 pixels for level 3).

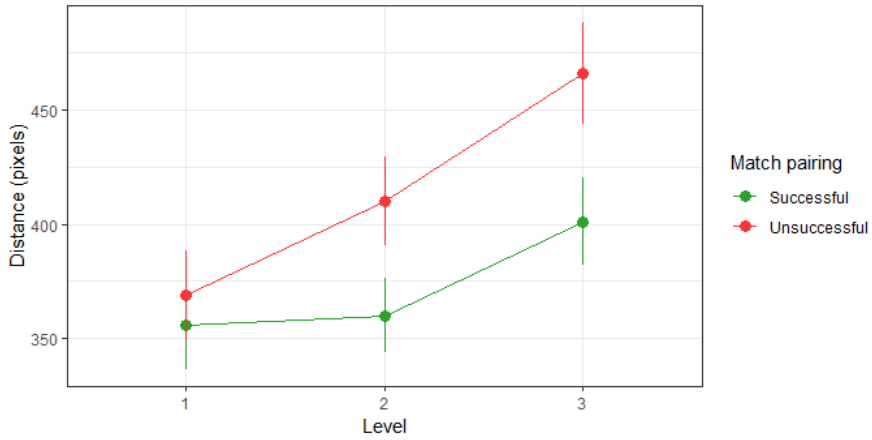


Figure 5.19: Median distance between gaze and stationary hand event positions.

When studying the results per participant, there is not a systematic pattern we can observe, even when the difference is significant on a per participant basis. In some cases, the results aforementioned in this section are not applicable to a participant (similar or greater distance for successful stationary hand events), as shown in Appendix C, Figure C.19.

5.6 Discussion

Our work contributes to the understanding of the gaze/hand correlation in the context of touch devices. We retrieved stationary hand events, not only to supplement existing work solely focused on taps (Chapter 4), but also because we considered this specific event to indicate the indecision users may face while interacting with a tablet. We found that the relationship between gaze and hand during the stationary stage of target selection is closely dependant on the target's location, and that users keep their hand closer to them in the vertical dimension while they preferably moved it in the horizontal dimension. When the users' hand lingers during interaction, they may point at a location that matches a "manual mental map" rather than directly pointing at the same location than gaze. Therefore, the role different screen sizes, orientations and target dimensions may play on the visuomotor mechanisms and on the construction of this map, could impact the location of the hand during stationary hand events, and in return influence the correlation between gaze and hand.

In the detection of stationary hand events, our results indicated that IVT is always providing poorer F1 scores in all scenarios, and can be ruled out when looking for a

compromised between precision and recall. IDT and IDTE are behaving similarly, the choice of the algorithm would depend on what the application should specifically retrieve: hovers, dwells or both. When one event type is to be retrieved, IDTE appeared to be the best algorithm to use. The algorithms have not only been chosen for the similarity of the motion behaviour between eyes and hands, but also as an answer for dynamic detection. The instant velocity or the displacement of the hand can be retrieved with low latency from the Leap Motion data, and stationary hand events detected with an expected poor computing cost (the algorithms are simple and do not require too much memory, nor other hardware specification requirements) and a reasonable lag (the longest time threshold for the best performing configuration described in Section 5.3.3 was 220 ms). The present work serves as a baseline and an initial generic detection of the stationary hand events, a foundation for two possible future studies. Firstly, it constitutes a reference for designing personalised algorithms based on self-adapting thresholds (machine learning). Since individual differences were clearly observed during our study, and partly suggested by difference of values between the validation and the test subsets' metrics, we suppose that personalised algorithms may be more efficient than the generic version we worked with. However, the downside of such technique is the training time required to make the algorithm efficient during interaction. Secondly, with either versions of the algorithms (generic or personalised), we can evaluate *when* stationary hand events are indicating hesitation during decision making, based on a new data collection. Ultimately, detecting stationary hand events that convey the users' hesitation serves the field of Human-Computer Interaction because it allows applications to assess when users experience difficulties (in their choices or in their interaction), and thus propose alternative answers.

Integrating intelligence in machines to decipher human cognitive clues is a challenge [63]. We aimed at finding how indecision can be inferred from the gaze and hand correlation. Approximating the decision making cognitive states {decisive/indecisive} by the success of the tile pair matching on seen tiles, we found that contrary to our expectations, the number and duration of fixations during hover cannot reveal indecision. However, we noticed that during hover, the hand is closer to the point of gaze when the user is decisive, and that the vertical component of this distance brings this closeness. Surely, better indicators for indecision should be used to get a more accurate estimation of the users' state of mind. Nevertheless, our approach enables a first coarse estimation that may serve as a basis for future intelligent systems.

As the tiles were shown facing down when they were not flipped or paired, we can assume that the tiles did not intrinsically play a role in the gaze movement: players did not perform a visual search through concrete pictures to flip the tiles when they browsed the screen. Instead we can consider the gaze movements are directly related to the mental map the players are involved with [5, 102]. However, the role of the revealed paired tiles may be interesting to query, since they became visual and spatial cues for the players to retrieve the tiles that have not yet been matched.

For our data analysis, we did not take into account personal differences despite being already acknowledged in gaze/hand correlation for taps (cf. Chapter 4). When observing the participants playing, we saw that some of them did not move the hand almost unless that for tapping on the tile, whereas some others were more likely to often browse the screen with their finger. The correlation between gaze and stationary hand events may therefore be predicted up the users' manual and visual behaviour categorisation (personal differences for indecisive vs. decisive groups were found in [125, 154]). If a deeper analysis can be obtained from this categorisation, it must then be taken into account towards implementing more intelligent systems that adapt their response to the user's cognitive state (for example, using virtual agents to assist the user's choice when indecision is detected, or store the candidate choices the users hesitated about, to propose them again in the scenario when the first choice the users selected has been discarded between the selection and the validation/confirmation process). This intelligence in machines meets the definition of Langley "An adaptive user interface is a software artifact that improves its ability to interact with a user by constructing a user model based on partial experience with that user." [122].

5.7 Conclusion

We have conducted a data collection that encompassed gaze and stationary hand events data, on a touch enabled tablet while playing a Memory Game. Our objectives were to propose a detection method based on the analogy between the hand and gaze patterns, and to understand how the hand and the eyes correlate before the taps were performed, at a particular event when the hand was in a standby position. We observed that the distance between gaze and hand depended on where the user looked on the tablet. This distance increased radially from the bottom centre-right of the screen, and it varied more importantly in the horizontal axis. This behaviour supports the concepts of energy-

efficiency of hand movements [4**?**] and the predominance of horizontal gaze movements [44, 70, 74].

We also wanted to estimate how the correlation can inform about the participants' cognitive process. We compared the gaze/hand relationship for stationary hand events leading to successful tile pair matching with those leading to unsuccessful tile pair matching to approximate the participants' indecision as an example of cognitive state. We found that the number and the length of fixations did not depend on the indecision, but that the distance between the finger and the eyes was larger when a decision has been taken with uncertainty. These results are in line with observation of body response and interaction pattern changes observed in a situation of indecision in other research works [124, 154, 157].

We endeavour to explore the correlation in a more detailed approach by understanding how it differs on the personal level. We suspect the personal hand motion and/or gaze behaviour to have an impact on the correlation, and thus should be an element to consider to the implementation of a finer detection method of the users' cognitive process stages and propose mechanisms to assist the user (adaptive system behaviour described by Langley [122])..

6

Correlation between Gaze and Hand in Motion

6.1 Introduction

Behavioural studies on hand and eye coordination often only measure simple movements triggered by stimuli [156]. Certainly, the context of interaction with computing devices in hand-gaze correlation studies ensures naturalness of the movement, since various applications these studies are based upon are more likely to reproduce by subjects in their ordinary use of the devices. However, so far, hand movements are either studied to understand the correlation between gaze and hand via the mouse (i.e. [123]) for desktop computers, or suffer from constraints (lack of naturalness, often because of abstract tasks [1, 16, 17, 30, 62, 198]) or limitations (near field detection [89, 147, 219], wearable sensors [215]) on tactile surfaces. When users interact with a tablet, the hand spends a significant amount of time "in the air" for what is generally considered, from the commonly deployed tablet's point of view, as idle interaction. However, gestures or stationary events of the hands in this area above the screen can actually inform a lot on the user's behaviour (cognitive activity, loss of interest with the tablet's content, frustration etc.). In this chapter, we describe the correlation between the hands and the eyes when the hands are neither tapping nor marking a stationary event. In other words, we want to analyse the correlation between the hand and gaze when the hand is moving between taps and/or stationary events (dwells and hovers). Thus, this chapter builds a bridge between the prior Chapters 4 and 5. The context of this chapter is a Memory Game, application chosen for the study of gaze-stationary hand events in Chapter 5. In relation with this

chapter, the Memory Game affords two advantages: due to the cognition process required to solve the game, the associated data collection contains enough occurrences of stationary hand events and taps, and it provokes hand movements from the participants across all the different areas of the tablet's screen in different directions and lengths. We present our results following a top-bottom scheme: describing the correlation between gaze and hand movement starting with the general condition and then against the different factors the study allowed us to analyse. We demonstrate that the correlation between gaze and hand motion is stronger in the horizontal dimension, that it is influenced by the difficulty of the task and the type of hand events the motion connects, and without surprise by individual differences.

6.2 Data Preparation

6.2.1 Context

The data collection context was the same as the one described in Chapter 5: we devised a game on a tablet (Microsoft Surface Pro 4) to generate cognition activity and hand movements, called Memory Game. The system also comprised an eye tracker (Tobii EyeX at 60 Hz) and a hand movement tracker (Leap Motion at approximately 110 Hz). Therefore, the data we collected were the eyeballs position in space and the estimated gaze points on the tablet, and the hand position in space. After a check to remove poor quality data (cf. Chapter 3.1.3), the dataset comprised elements from 71 participants (in total 177 game rounds, each participant played 3 rounds of increasing level). Further details on the apparatus and the participants can be found in Chapter 5.

6.2.2 Method

In order to estimate the correlation between gaze and hand during hand movements, we needed to align the data samples correctly.

From the eye tracker, we retrieved, for each round, the following information:

- timestamp,
- gaze sample position on the tablet (X and Y).

From the hand tracker, we retrieved, for each round, the following information:

- timestamp,
- palm and finger tips position in space (for each hand).

For the hand position, we only kept a single point in space as explained in Chapter 5, Section 5.3.1.2 (closest tip from the tablet surface).

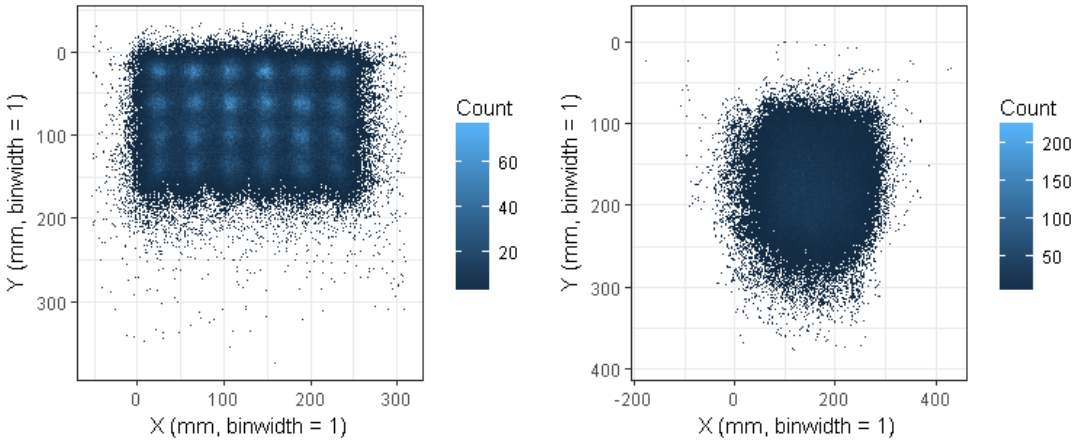
As we focused on the correlation between gaze and hand **during hand movement**, we only extracted the samples of the overall dataset that corresponded to the participants' hand movements. If the data occurred during a stationary hand event or around a tap (in a 100 ms time window centred on the tap), it has been discarded (stationary hand events were detected as mentioned in Chapter 5 and taps were already recorded at the tablet level), resulting in chunks of hand movement data.

In order to compare the gaze position samples with the hand position samples, two more steps were necessary. First, we computed the *projection* of the hand on the tablet, based on the eyes position (cf. Chapter 5, Section 5.3.1.2, Figure 5.4). Then, we aligned the data in time. Since the eye tracker generated data at lower frequency than the hand tracker, we *interpolated* the hand position at the timestamps provided by the eye tracker (with the estimation of a linear relationship between time and space between two consequent hand positions, coherent with Fisk and Goodale's estimation of limb movements approximating a straight line path when reaching targets [64]).

We finally generated an aggregated dataset containing both gaze and interpolated hand projection positions. In order to discard any data that were not strictly related to the game playing sessions (i.e. after finishing a round participants still stayed visible to the sensors without playing anymore), we retrieved the gaze and hand data only between the first and the last taps (first and last flipped tiles) of the game round.

6.3 Results

After preparing the dataset as mentioned in the previous section, our working dataset consists of 600,606 pairs of gaze and hand projection points, matching a total of 24,517 hand motions.



(a) Gaze. (b) Hand.

Figure 6.1: 2D distribution of the gaze and hand data.

6.3.1 Overall Correlation

We study the correlation between gaze and hand when the hand was moving in the air (which excludes tapping). The distribution of point coordinates for each modality and per axis is illustrated by Figure 6.1. It shows that gaze (Figure 6.1a) relatively to the *all* screen area is not normally distributed: the concentrations of highest values that appear on the plots reveal that participants principally points at the centre of the game's tiles. During hand motion, gaze follows a multimodal distribution corresponding to the game's structure.

As a consequence of at least gaze data not meeting a bivariate normal distribution we use Spearman's rank order correlation rather than Pearson's correlation to compute the correlation coefficient between gaze and hand coordinates for each dimension of the screen. On the horizontal axis (X), we find a correlation coefficient ρ=0.69 ($p<0.01$); whereas on the vertical axis (Y) we find a correlation coefficient ρ=0.58 ($p<0.01$).

Therefore, in the general situation, gaze and hand movements are weakly correlated in space. The correlation is nevertheless stronger on the horizontal dimension than on the vertical dimension. This result is coherent with the observations we made regarding the correlation between gaze and taps (in Chapter 4) and stationary hand events (in Chapter 5). After modifying the correlation coefficients to comparable statistical information with a Fisher Z-transformation, we compare the coefficients between X and Y axis: the correlation between gaze and hand movements on the horizontal axis is significantly higher than the correlation between gaze and hand movement on the vertical axis (Steiger tests

for dependant samples[1], Z=102.65, p<0.01). Maybe this stronger alignment on the horizontal axis can be explained by two causes: (1) the human system favours horizontal yaw movements over vertical pitch movements (better horizontal smooth pursuit [44], horizontal movements involve less muscle in action (medial and lateral rectus muscles) than the vertical movements (superior and inferior rectus muscles as well as the oblique muscles), saccadic movements in natural visualisation contains more horizontal movements [70, 74], visual field wider in the horizontal dimension [205]) and (2) moving the hand horizontally is preferred by users [118].

6.3.2 Correlation per Participant

When observing the participants during the data collection, we clearly noticed that some of them had a stronger tendency than others to keep their hand in motion between taps and/or stationary hand events. Thus, we want to understand how personal difference influences the overall results of Section 6.3.1, and if a categorisation can be observed.

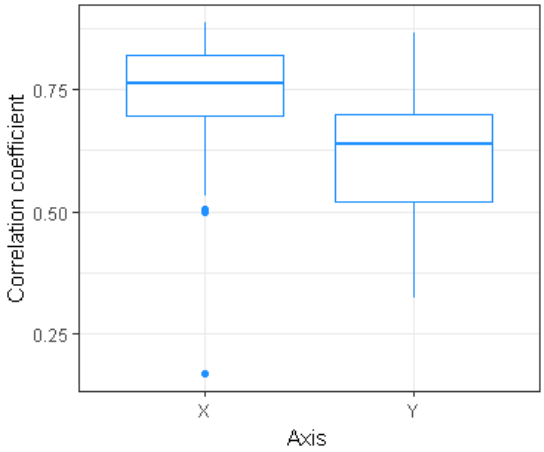


Figure 6.2: Participant's Spearman correlation coefficients boxplots per axis.

Figure 6.2 shows the boxplots for the Spearman's correlation coefficients of each participant on each axis (X and Y). For all participants, all the coefficients between X and Y are significantly different (p<0.01).

We still observe higher correlation coefficients in the horizontal dimension than in the vertical dimension, but clearly the correlation coefficients vary among users. In the horizontal dimension, they span from 0.17 to 0.89 and in the vertical dimension, they span from 0.32 to 0.87, as shown in Figure 6.3. However, the distribution of the correlation coefficients on the X axis among participants clearly indicates a smaller variance than

[1] As mentioned in `http://quantpsy.org/corrtest/corrtest3.htm` (last accessed Jan. 2020)

on the Y axis. We did not observe any apparent correlation between the percentage of time the participants' hand was in motion (measured between the first and last taps of each round, mean 66.5±7.9 %) and the strength of the correlations mentioned above.

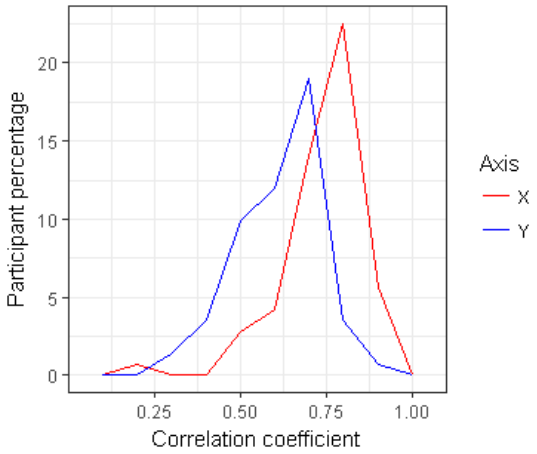


Figure 6.3: Distribution of the Spearman correlation coefficients over participants for X and Y axis (all levels).

We have shown in the previous section that generally, the correlation between gaze and hand movements is stronger on the horizontal dimension than on the vertical dimension. Figure 6.4 illustrates the Spearman's correlation coefficients relationship between X and Y for both dimensions per participant.

We clearly realise from Figure 6.4 that most of the X-Y correlation coefficients pairs are situated below the equality diagonal (where $Y = X$), illustrating the statement that alignment is stronger horizontally than vertically. Nevertheless, from this plot, we cannot identify distinct clusters that would divulge a classification of the participants, the correlation coefficients are all over the boundaries mentioned above.

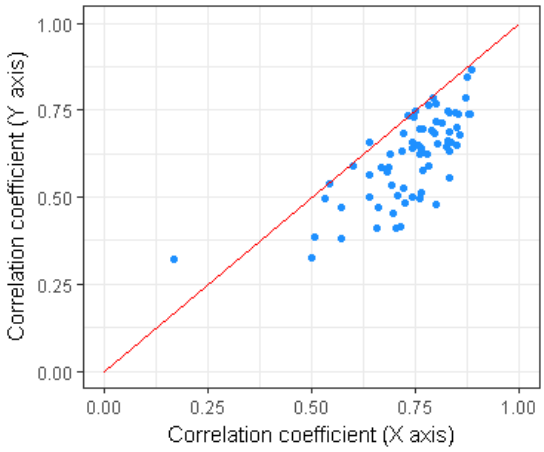


Figure 6.4: Spearman's correlation coefficients relationship over the participants for X and Y axis (all levels).

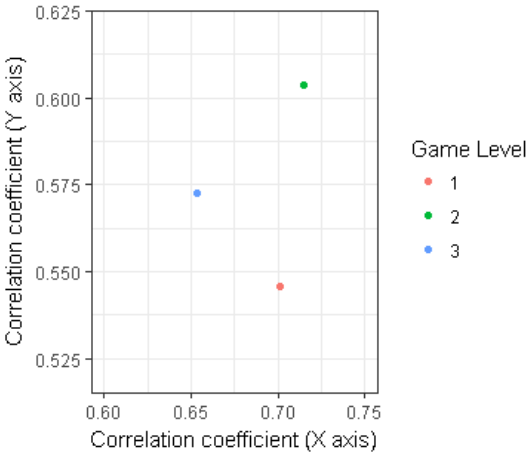


Figure 6.5: Spearman's correlation coefficients relationship over the game levels for X and Y axis.

6.3.3 Correlation per Game Level

Since we designed the game to trigger more intense cognitive mechanisms with the progression of the levels, we wonder whether the task difficulty influences the correlation between gaze and hand during hand motion. Correlation coefficients of each level in both horizontal and vertical dimensions are plotted in Figure 6.5. We compare each correlation pair-wise (after applying a Fisher Z-transformation, summarised in Table 6.1). Although the Spearman correlation coefficients difference between each level's pair is significant, we cannot observe a systematic increase or decrease of the coefficient values with the level difficulty. Nevertheless, the following remarks can be raised. For the X axis, the correlation coefficient for the level 3 (ρ=0.65) is lower than the two others (ρ=0.70 for level 1 and ρ=0.72 for level 2), suggesting that when the task difficulty increases, users tend to weaker align gaze and hand horizontally. On the Y axis, even if the correlation is always weak, we observe that the coefficient value (ρ=0.55) is worse for the easiest level than for the two others (ρ=0.60 for level 2 and ρ=0.57 for level 3), indicating that when difficulty is low, vertically, users do not align gaze and hand as much.

Table 6.1: Pair-wise Z-score for the Spearman correlation coefficients between gaze and hand comparison per game level.

Level Pair	Z-score (X axis)	Z-score (Y axis)
1-2	8.23*	26.12*
2-3	38.5*	15.7*
1-3	27.41*	12.17*

$^{(*)}p<0.01$

In sum, the correlation between gaze and hand on the X and Y values seems to be influenced by the difficulty of the task in opposite ways: weaker when the difficulty is

higher for the horizontal dimension, and weaker when the difficulty is lower for the vertical dimension. This behaviour could be explained by more demanding cognitive efforts from the users which leads to more vertical gaze movements than in normal situation (this behaviour has been observed in another context, driving [165]).

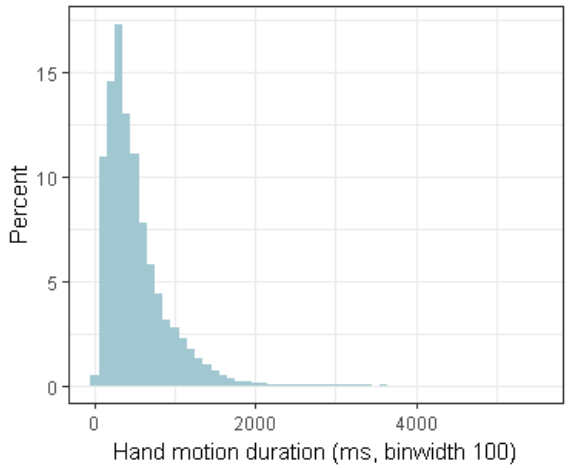


Figure 6.6: Hand motion duration distribution.

6.3.4 Impact of Time on the Correlation

Each hand movement between taps and stationary hand events varied in time from milliseconds (84 ms) to seconds (5.49 s), as shown in Figure 6.6.

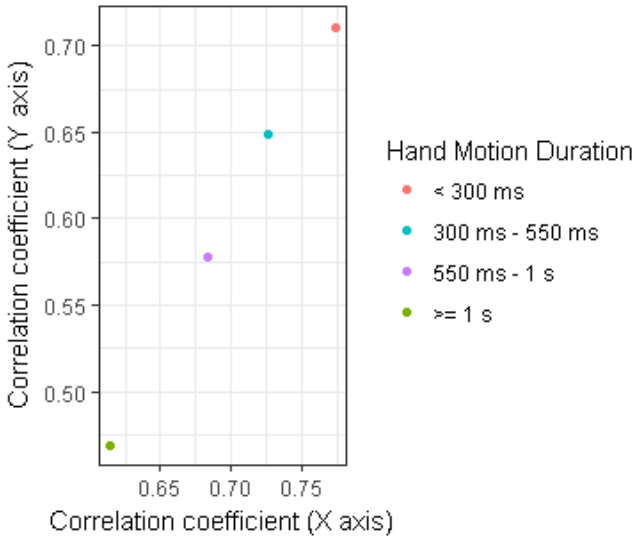


Figure 6.7: Spearman's correlation coefficients relationship over the motion duration ranges for X and Y axis.

We investigate the role played by the duration of the hand motion on the correlation between gaze and hand. To do so, we categorise the data into different time ranges,

based on the approximation of the quartiles' value: `less than 300 ms`, `between 300 ms and 550 ms`, `between 550 ms and 1 s` and `more or equal to 1 s`. Figure 6.7 illustrates the correlation between gaze and hand during hand motion for each time ranges mentioned above.

Table 6.2: Pair-wise Z-scores for the Spearman correlation coefficients between gaze and hand comparison (X axis) per duration range.

Duration range	<300 ms	300 ms - 550 ms	550 ms - 1 s	>= 1 s
<300 ms		25.15*	46.34*	72.76*
300 ms - 550 ms			25.46*	58.38*
550 ms - 1 s				35.6*

$^{(*)}p<0.01$

After executing a Fischer transformation, we compare the different correlation coefficients pairwise. The Z-scores are reported in Tables 6.2 and 6.3 (for the X and Y axis respectively, $p<0.01$ for all pair comparisons). Figure 6.7 suggests a trend for the correlation to be degraded when the duration of the movement lasts longer (on both the horizontal and vertical dimensions). Indeed, the correlation score is very poor for the group of hand movements lasting more than 1 s.

Table 6.3: Pair-wise Z-scores for the Spearman correlation coefficients between gaze and hand comparison (Y axis) per duration range.

Duration range	<300 ms	300 ms - 550 ms	550 ms - 1 s	>= 1 s
<300 ms		26.98*	54.86*	88.28*
300 ms - 550 ms			33.72*	75.17*
550 ms - 1 s				44.9*

$^{(*)}p<0.01$

Results in Chapter 5.5 showed that the distance between gaze and stationary hand events is wider when the participants were considered as indecisive. We suppose that the hand movements preceding the stationary hand event could be poorly correlated with gaze and, in a situation of indecision, therefore may require more time to be performed (as longer movements are less correlated with gaze).

6.3.5 Impact of the Motion Type

The hand movement above the tablet's screen may carry different information depending on where it starts from and ends to. For example, when the motion is strictly above the tablet's surface, we can infer the user is either in a decisive action (when movements are

short in time) or in a search action (when movements take longer). A movement that ends outside the volume strictly above the tablet's surface may indicate the user is in a phase of observation (the hands retracts from the interaction space to give a full visibility of the interface) or even indecision (loss of engagement) [3, 71, 175, 201].

We distinguish 9 types of motion to cover all the combinations between taps, hovers and dwells (the two stationary hand events described in Chapter 5, Section 5.3.1.2):

- `T-T` from a tap to another tap,
- `T-H` from a tap to a hover event,
- `T-D` from a tap to a dwell event,
- `H-H` from a hover event to another hover event,
- `H-D` from a hover event to a dwell event,
- `H-T` from a hover event to a tap,
- `D-H` from a dwell event to a hover event,
- `D-D` from a dwell event to another dwell event,
- `D-T` from a dwell event to a tap,

The distribution of the hand motion types is illustrated in Figure 6.8. It appears that most movements are of type `T-T`, and that more than one stationary hand event between two taps occurs rarely (fewer motion connecting two stationary hand events).

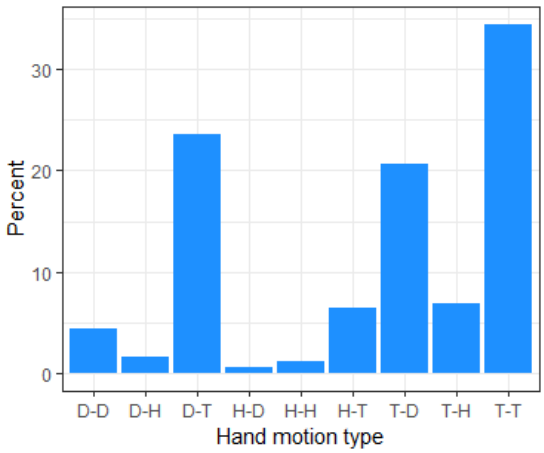


Figure 6.8: Hand motion type distribution.

We compute the Spearman correlation of each type between gaze and hand during hand movement on the horizontal and vertical axis. Figure 6.9 represents each correlation coefficients for both axes. As observed in other cases, the correlation between gaze and hand during hand motion is more important horizontally than vertically, for all types

of movements. However, two clear clusters of points on the graph suggest that the correlation is much stronger when a tap is involved into the hand motion (either at the beginning or at the end of the movement).

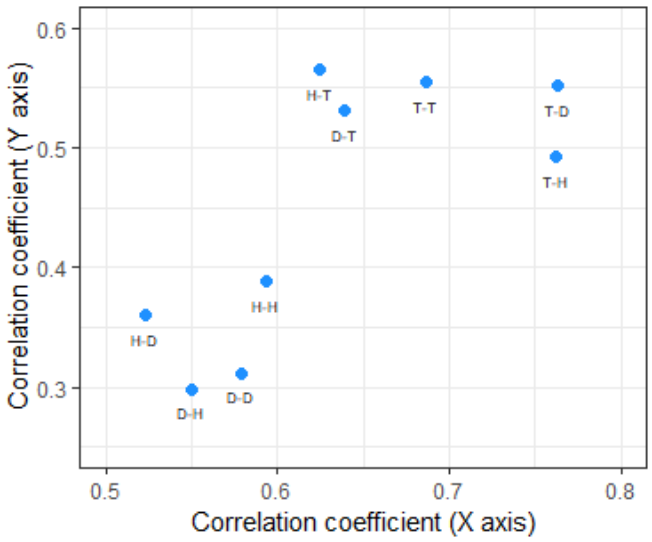


Figure 6.9: Spearman's correlation coefficients relationship over the motion types for X and Y axis.

By comparing the correlation coefficients pairwise using a Fischer transformation, we find that all the correlation coefficients for the motion types involving a tap are significantly stronger than those only involving a dwell or a hover ($p<0.01$ for all comparison, results summarised in Appendix D, Table D.1 for the X axis and Table D.2 for the Y axis).

6.3.6 Spatial Difference between Gaze and Hand

The context of the study being a Memory Game, we take the game's tiles as a reference for the spatial division of the tablet's screen (each tile is a square of 304 pixels/43.38 mm side). Figure 6.10 illustrates the mean difference between gaze and hand position during hand motion on each tile (where gaze is located in a tile).

The difference between gaze and hand position during hand motion changes across the tiles: we observe a closer distance between gaze and the projection of the hand in the bottom right of the tablet screen, that increases radially toward the top corners. The horizontal difference is, once again, showing a closer alignment between gaze and hand across the screen's width (span of 81.19 mm, 31% of the screen width) than vertically (span of 40.31 mm, 23 % of the screen height). Generally, the hand tends to keep close to the body, hence the radial increase of the distance. Horizontally, results indicate that the projection of the hand crosses the line of sight somewhere close to the right

part of the tablet, supporting the idea that users limit their hand movements (result indicated by the fact that most participants were right-handed as mentioned in Section 5.2.5). Vertically however, the hand's projection always remains below the line of sight. The relationship between gaze and hand during movements followed the same principles exposed in Chapter 5 between gaze and stationary hand events: the hand tends to move at a minimum effort, hence favouring horizontal movements. The distribution of the gaze/hand difference over the screen informs on the mental maps the users generate while the hand stays in the volume above the tablet. Nevertheless, these results can only apply to a system, such as a public kiosk, that does not require scrolling (in particular, vertical scrolling). Results observed for such interaction may differ as it bring other hand gesture that were not required when playing the Memory Game.

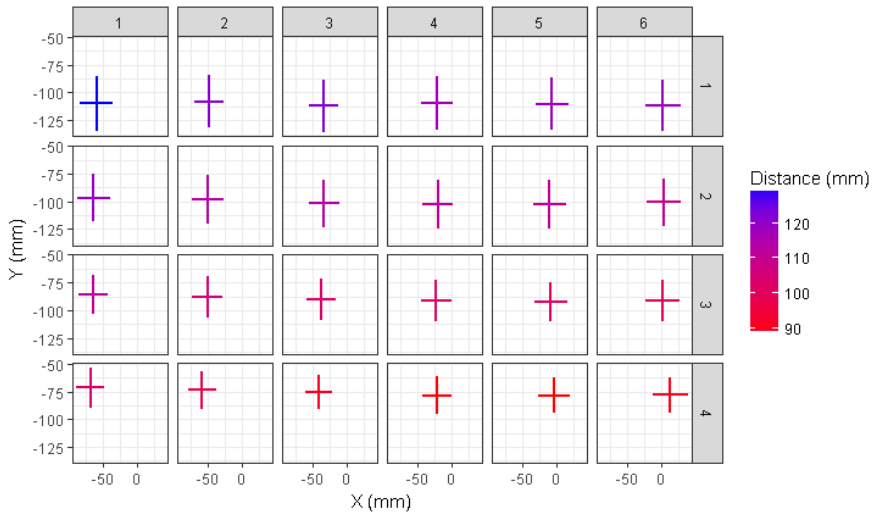


Figure 6.10: Mean difference between gaze and hand during hand motion per tile (based on gaze location).

6.4 Discussion

The correlation between gaze and hand during hand motion bears similarities with the observations made on the relationship limited at taps or stationary hand events: gaze and hand align stronger on the horizontal dimension than on the vertical dimension. Beyond the expected individual differences that resulted in various correlation strengths, we have also explored how several characteristics inherent to the application context (level of difficulty) or to the motion (duration and type) influence the relationship between gaze and hand while the hand was moving. The results we found can be explained in the following ways. We observed that the difficulty level of the application result in a weaker

correlation on the horizontal axis and a stronger correlation on the vertical axis when the difficulty increases. We suppose that the reason for this behaviour is a tendency for the users to keep their finger aside the tablet when the difficulty is higher, resulting in a movement mainly following the vertical axis rather than skimming the screen in both dimensions (maybe as a consequence of unconscious willingness to keep the visual field free from the hand and to place the hand as a marker of a potential selection to perform later). Guessing the perceived difficulty of an application can therefore be achieved by observing how the correlation between gaze and hand during hand movement behaves.

Based on the motion's duration and type (connection between the different events, taps or dwells or hovers), we found that the correlation is stronger (both for the X axis and the Y axis) on shorter movements, and when the motion connects at least one tap with another event. The degradation of the correlation over time may be understood as a consequence of heavy cognitive process: when the hand movement lasts, chances the user is involved into a more intense reflection are higher, resulting the hand to probably "wander" while the eyes keep searching the mental representation the user has of the application. The results found with the type of motion also support this explanation, since a movement between two stationary hand events is most likely to be associated with hesitation or cognitive load task.

Our work analysed the role cognitive load tasks play in tablet interaction based on three increasing difficulty levels of a game. This low number of levels and the complexity of rating the degree of difficulty may be pointed out. However, we are confident in our choice for most participants had perceived an increasing difficulty in the levels (a feeling that they expressed during the data collection - unfortunately we did not collect their feedback over the level perception), but we are fully aware that more detailed and rigorous evaluation can certainly be performed with tasks containing more increments of difficulty, and a reliable evaluation scale associated with the difficulty of the levels (for instance, Paas et al. proposed measurement means associated with the Cognitive Load Theory [150]; Brünken et al. defined a framework in a multimedia based context [23]).

The results we expressed are valid for tablet interaction that did not require scrolling. If the tablet were used for browsing the Internet, or using social media, our observations may differ greatly because of the scrolling that generates hand movements are not correlated to gaze movements [193].

The data collection we devised allows the study of predictive hand movements duration with the expression of the Fitts' law [65]. Fitts proposed the expression of the *index*

of difficulty (ID) as a function of the distance to the centre of a target (D) and the width of this target (W), and also expressed the *index of performance* (IP) as a function of the index of difficulty (ID) and the movement time (MT). His original expressions have later been adapted to Human-Computer Interaction (MacKenzie [126]) Therefore, different movement times between two tiles of the game may inform how much cognitive load the participants are experiencing [78, 127]. Nevertheless, Fitts' law models well movements that are clearly targeted (aim reaching tasks) and without a baseline, a system that estimates the cognitive state of a user by exploiting Fitts' law may not be able to perform well before a training set is available to compare actual data with.

6.5 Conclusion

In this chapter, we have analysed the correlation between gaze and hand when the hand was in motion, between consecutive tap, hover or dwell events. The context of the data collection, a Memory Game, allowed the participants to generate enough manual interaction on the tablet and was a good example of an application which required a cognition activity, with an increment of difficulty over the three levels, making the hand moving between taps and stationary events throughout the volume above the whole surface of the tablet's screen.

We studied the relationship between gaze and hand motion spatially, in a general situation and according to several factors inherent to our study. In general, we found that the correlation between gaze and hand during hand motions was good in the horizontal axis (ρ=0.69) and average on the Y axis (ρ=0.58). A stronger correlation on the horizontal dimension is in line with the results found in previous chapters. When studying the relationship between gaze and hand during hand movements according to different criteria, we also found that a stronger horizontal alignment was maintained. These results are coherent with previous findings related to stationary hand events and gaze correlation (Chapter 5), and should support an "horizontal-layout" design of tablet's applications to increase the user's experience on those devices.

The first criteria we observed was the individual factor. It was expected that the correlation would vary among users. The correlation was still stronger on the X axis for most users, and the individual differences were more important vertically than horizontally (the correlation varied more for the vertical axis). We associate this result with personal choices of limiting hand efforts or not.

We also studied the correlation according to the task difficulty, the duration of the motion and the events the motion connected (type of motion). We showed that considering all factors (longer movement, more than one stationary hand event and task difficulty), the cognitive load impacted the correlation in such ways: the more the complexity the less eyes and hand correlated, in both dimensions (except vertically where the correlation was slightly better, while poor, when the difficulty increased).

These findings are elements that can be considered in the design of intelligent devices that adapt to the user's behaviour, for example detecting difficulty of interpreting the application's content from user's changes in the way her gaze and hand movements are correlated during interaction.

7

Discussion

This thesis presented the correlation between gaze and hand at different stages of the hand activity found during interaction with tablets. Research focusing on human behaviour often leads to an exploratory phase: would the finding be representative of most persons, or would several profiles emerge from the analysis of the data? In our case, we expected individual differences as we observed the participants during their interaction. Experience with tablets, physical/cognitive abilities and emotional state form a non exhaustive list of some speculative factors that may explain why people do not interact in the exact same way. Despite this expectation, we tried to draw conclusions from general data that describe an *average* behaviour, while acknowledging the differences. Therefore, deeper studying the individual differences is a potential continuation of the work related in our thesis, such as investigating whether the correlation between gaze and hand during tablet interaction *systematically* depends on specific behaviour categorisation (of gaze patterns, hand patterns or any other modality). In this chapter, we address the limitations of our research work we retrospectively analysed, as well further research ideas that can derive from the work we have presented.

7.1 Limitations

7.1.1 Apparatus

For both data collections we covered in the thesis, we used tablets that had similar dimensions and that were always presented to the user in a landscape orientation, because we wanted a setup that is similar to a public kiosk (that are mainly in this orientation) . As we have tailored our research work to be reflecting the users' *natural* interaction, the tablet orientation could be an element that greatly impact the results: because some users may prefer interacting with the tablet in the portrait orientation, and because the visual dynamics can change greatly (the change of orientation of pictures influences the gaze directions [70]). Tablet's position is another variable that we locked. Even if the apparatus differed between each data collection (Tobii rack for the first, home-made support for the second), the tablet laid in a near to horizontal position. Although this position is used for public kiosks (i.e. in a museum) and for personal interaction by users, vertical (or near to vertical) displays are very common (i.e. in ticketing machines and public displays). Our methods to explore the correlation between gaze and hand would not changed, but the orientation and/or position would be new variables to take into consideration.

Eye tracking technology has certainly become more reliable and efficient. Nevertheless, remote eye tracking based on infrared corneal reflection still suffers limitations: light needs to be controlled, interferences need to be avoided, and the subjects in front of the eye tracker may not comply with the physiological requirements necessary for a good tracking. Hence, studies involving eye tracking are, by nature, limiting the population on which analysis are based upon (which again accounts for the complexity of working on human factors). If applications can be devised from our results, it will concretely be difficult to implement them in the real world: there is no guarantee to meet the controlled environments of the laboratories once crossing their doorsteps. Future work based on our results will most likely have to face these challenges, by choosing tracking tools that are not sensitive to external stimuli, yet not obtrusive if naturalness needs to be maintained.

7.1.2 Tasks

Throughout the work associated with this thesis, we have favoured a study context that reflected as much as possible ordinary activities that can be performed by users

on tablets in order to capture their natural behaviours: Internet search, shopping and link following on Wikipedia (Chapter 4) and a game (Chapters 5 and 6). Necessarily, especially for the data collection context presented in Chapter 4, this approach led to some difficulties in the analysis. Since there was no predefined targets to steer taps towards, the content displayed by the tablet (for Chapter 4) and the hand movements (for all chapters) were totally random: they were entirely dependent on the participants' choices. More often, abstract tasks are performed by participants in studies that research on hand and gaze correlation, possibly because it provides results for the very elemental behaviours observed in humans. With a more interactive context, distractors are impacting the attention of the participants and probably influence the way the gaze at the scene. Still, the data collection contexts we selected were more practical than what presented in other studies related to gaze and hand coordination. In addition, even if we had chosen to work with abstract tasks, it would have just been good enough to prepare a baseline work, on which we would have certainly required to build on with the same study contexts we eventually presented in this thesis to elaborate how factors occasioned by Human-Computer Interaction alter the correlation between gaze and hand.

We studied in-the-air hand interaction (stationary hand events in Chapter 5 and hand movements in Chapter 6) based on a data collection's context that did not require the user to scroll or zoom, whereas the data collection's context we relied on to study the correlation between gaze and taps (Chapter 5) allowed the users to scroll or zoom, but was not be recorded with our setup. Therefore, we strongly recommend further research based on our work to combine the apparatus and data collection's context in a way that these different gestures are triggered (application that requires scrolling and zooming) and measured (with a hand movement sensor or from a tablet-based gesture recognition tool).

In Chapter 5, we have estimated the hesitation of the participants based on the wrong pair matching instances of already seen tiles. More reliable work on hesitation would have been achieved if the detection of the nonsystematic hand gestures and tasks that can clearly scale the difficulty levels had been in place to define the ground truth to compare our data with. Standard evaluation scales must preferably be used to evaluate the difficulty of a task in a much objective angle as possible. Besides a clear infrastructure framework to evaluation difficulty or hesitation, further body movements recording and analysis could have been considered, such as shoulder shrugging, frowning etc. Our data collection setup did not allow such recording.

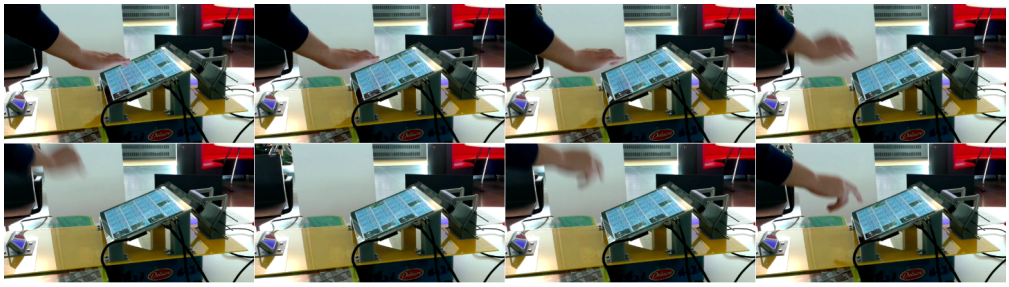


Figure 7.1: Example of the hand leaving the interaction space as a potential expression of frustration, hesitation or reflection.

7.1.3 Data

Although our participants' recruitment did not discard people, our population is representative of healthy individuals with none to moderate eye correction. As a consequence, so is our data, and the results we obtained may not be valid for subjects suffering from major visual impairment, physical or mental disabilities.

We considered the hand events to structure the thesis upon (taps, stationary hand events and motion in between) as the common and basic hand events encountered during interaction with a touch enabled device, and therefore, we limited our work on those events to cover the essential of the hand activity over a tablet. Nevertheless, we are aware that many other more complex gestures are performed by users such as a flip of the hand to express surprise or frustration, small repetitive movements of the fingers to express impatience or reflection, or more drastically when the hand is leaving the interaction surface and acts on the body to express a doubt (as illustrated by Figure 7.1). These other gestures are sometimes specific to an individual (nervous tics), thus tackling the gestures puts research on the edge of generalisation when dealing with human factors. These individual gestures have been ignored in our work, which necessarily has an impact on the study dealing with hand motion (Chapter 6): the hand motion connecting taps and/or stationary hand events were not always continuous nor constantly followed a "direct" path.

7.2 Further Research

7.2.1 Human Factors and Sensors

We have considered only specific factors that can impact the users' interaction with tablets: individual interaction patterns in typing and estimation of indecisiveness. These

factors are far from being exhaustive, and we suggest several other factors than may complete our work and provide research with further insights on Human-Computer Interaction with tablets. Human-Computer Interaction has benefited greatly from new sensors that have placed the hand as the "old" modality. This thesis treated the correlation between gaze and hand because they are the two main body parts involved in the relationship between humans and touch enabled devices. Human body language is far more informative and complex than the taps which computers commonly and solely interpret at this stage. Therefore, other body measurements that indicate the users cognitive or emotive state (such as shoulder shrugging [111], heart rate and blood pressure [90], facial expressions [60]), already sometimes monitored by wearable devices, may soon enough be consistently integrated in the data processed by computers to understand their users and respond adequately, shaping the reality of Intelligent Human-Computer Interaction.

- **Frustration** As mentioned in the limitations (Section 7.1), users may show signs of frustration, impatience etc. We invite researchers to classify hand gestures to detect and take frustration into consideration in further work via running gesture detection algorithms on hand movement sensing data or analyse video recordings (computer vision or annotation). Ad hoc detection is possible if the hand gestures that are associated with frustration are already listed and known before the data collection. Otherwise, post hoc data analysis needs to be run. We suggest to use another sensor than Leap Motion to detect the hands, if a sensor is preferred to video recording: Leap Motion may not perform well in tracking hand movements that are far away from the tablet (or the device's field of view) and occlusion may prevent from reliable sensing. Depending on the context of the study, wearable devices may be considered i.e. wearable accelerometers [218]. Frustration can also be detected from other body response like frowning and audio recording.
- **Engagement** Engagement is a reliable criteria to detect how users are responding to an application in Human-Computer Interaction studies [145]. To detect it, our data collection can be sufficient: gaze and hand movements data can be analysed to detect loss of attention and engagement. However, validation may be required by further video analysis (i.e. when the sensors do not track movement, does it mean gaze or hand has left the field of view corresponding to the interaction area or is it a sensor issue?) and therefore, considering a video recording of at the least the upper half of the body may be necessary. Complementary detection can be

encompassed by analysing audio recordings of think aloud scenarios. Of course, system inactivity can also reveal lack of engagement, but should be carefully used to avoid false assumptions (for example considering the completion of a task as inactivity that assumes a lack of engagement).

- **Arousal** Arousal has been considered in Human-Computer Interaction as a factor that can be monitored by intelligent systems to enhance their response to users' emotions [137]. To achieve so, different sensors need to be used, such as skin conductivity sensor [85], Electroencephalograph, heart beat rate sensor, electrodermal activity sensor, respiration sensor, finger thermometer or facial expression analysis [137]. Most of these sensors break the naturalness of the interaction we wanted to preserve by using remote devices. However wearable devices can be a solution to avoid the users' feeling of obtrusion [163], such as the Empatica E4 wristband[1]. The sensors we used in our study could potentially be used by measuring the pupil dilation of the users, however, pupil dilation can be triggered by external factors (such as the change in ambient light) and is not reliable on its own.
- **Anxiety** Computer interaction is known to be a cause of anxiety [86, 212]. Even if suppressing anxiety may be very challenging, detecting under which conditions peaks of anxiety are found during interaction with touch-enabled devices may be a research area to tackle, so that intelligent systems decide which scenarios to adopt when anxiety is detected. This factor cannot be measured with the methods we used during our work. Anxiety is often measured via the skin conductance rate (sweat levels) and heart bit rate. Moreover, further research towards evaluating anxiety must allow a data collection in two steps: one that first serves as a baseline for when subjects are in a "relaxed" situation (i.e. listening to music) , and another one that actually focus on the interaction with the touch-enabled device.

For all factors, video recording of a larger portion of the body (at least the full upper half of the body to cover the torso, the limbs, shoulders, and the face) can to be used, either for ad hoc data logging (for example using computer vision) or post hoc video analysis.

[1]Empatica E4 EDA/GSR sensor webpage `https://www.empatica.com/research/e4/` (last accessed Jan. 2020)

7.2.2 Contexts

As we thought of public kiosks as a target for our work, we did not consider the way users *hold* the tablet. In our work, the tablet could not be moved and carried as it is in public spaces. If similar work than ours was to be considered for *personal* use of the devices, the different holding strategies deployed by the users can be explored (for instance, are the results changing when only one hand is used for interaction while the other hand holds the tablet, compared to users who leave the tablet on a surface and interact with both hands).

We have tried to find data collections' contexts that keep the naturalness of the interaction with a tablet as much as possible. However, since we have thought our work to be relevant for applications in public spaces (such as kiosks), different contexts could have been explored to also reflect public spaces' applications (i.e. ticket booking/collection, map reading, browsing). We also deliberately ignore tablet's professional usage (such as productivity software interaction) to focus on widespread activities (according to a 2018 Ofcom report[2] tablet users spent twice more time online with their tablet at home and at any location than at work; Müller et al. surveyed tablet's users in 2011 and concluded that "*[...] the most frequent activities [are] checking emails, playing games, social networking, and looking up information. [...] tablets are primarily used for personal purposes[...]*" [138]).

As discussed in the limitations (Section 7.1), different configuration of the tablet, as the orientation and the position, can be explored to better understand the correlation between gaze and touch other scenarios. We also encourage further research to address the correlation between gaze and hand for application on *mobile smartphones*, as they represent the large majority of touch-enabled devices used wordwide (3.3 billions smartphone users 2019 forecast according to Statistica[3] in 2019 versus 1.23 billions tablet users 2019 forecast[4]).

We strongly recommend further research to evaluate new systems based on the work we presented, in order to test applications in a real context.

[2]Ofcom Communications Market Report 2018 `https://www.ofcom.org.uk/__data/assets/pdf_file/0022/117256/CMR-2018-narrative-report.pdf` (last accessed Jan. 2020)

[3]Number of smartphone users worldwide from 2016 to 2021, Statistica `https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/` (last accessed Jan. 2020)

[4]Number of tablet users worldwide from 2013 to 2021, Statistica `https://www.statista.com/statistics/377977/tablet-users-worldwide-forecast/` (last accessed Jan. 2020)

7.2.3 Replication

We designed two data collections that can be partly (for the first one) or entirely (for the second one) reproduced by fellow researchers. The apparatus in both data collections can be reproduced: the tablets, eye trackers, hand tracker are commercial devices, so as the rack used to maintain the tablet and the eye tracker in place in the first data collection (cf. Chapter 4). The apparatus support board we designed for the second data collection (related to Chapters 5 and 6) can be reproduced: we provide the dimensions required to dispose the apparatus' elements (tablet, Leap Motion and eye tracker) relatively to each other (Appendix B, Figures B.1 and B.2). Only the context of the second data collection, however, can be reproduced totally as it is a game that has a static presentation. The first data collection's context can be partly reproduce: questions of the first task, rules for the game tasks and shopping activity for the shopping task. Not only the underlying Internet content will vary among users (as we reported) but it may also differ greatly from when we did our data collection, for example if the shopping website workflow changed dramatically.

As mentioned in Section 7.1, we advise potential replication of our data collections to deviate from the contexts/apparatus we designed so that the context triggers richer gestures such as scrolling or zooming, and the system effectively tracks them (either by a tablet-based tool or by the hand movement sensor). Based on our work, tracking more gestures would require to modify Sparsh-UI to detect them (cf. Section 3.2 in the case the first data collection is reproduced, and allow gesture detection with Leap Motion if the second data collection is reproduced.

7.2.4 Applications

We orientated our research work towards finding how the relationship between gaze and hand during the interaction with a tablet can be used to infer the user's indecision. This is one example of user's emotions or cognitive processes that we have chosen. We discuss in this section suggestions on how adaptive systems should implement a detection to user's indecision (and other emotions/cognitive processes) to meet with the principle of Intelligent Human-Computer Interaction proposed by Pantic et al.: "'*Human-Centred Intelligent HCI (HCI²) must have the ability to detect subtleties of and changes in the user's communicative behaviour (as expressed through, e.g. affective and social signals), and to initiate interactions based on this information, rather than simply responding to*

the user's commands." [152].

Hesitation in computer interaction can occur during an expected selection process (such as the choice of a product) or during an unexpected interruption in user experience when the context the user interacts with appears to be unclear or ambiguous (design issue, labelling issue etc.). Therefore, intelligent systems which monitors the user's indecision must also be able to evaluate the right situation (expected or unexpected), as well as the causes of hesitation. In the first case, possible intelligent system's responses are: (1) displaying an assistant tool to help the user with selecting the right item, in a noticeable way that is yet unobtrusive, (2) when the interaction process implicates the validation of a selection, proposing the *second* choice the user did not make if the validation is cancelled (the user changed her mind). In the second case (unexpected situation), intelligent systems cant hint indications on how to use the application and/or self-adapt to prevent the ambiguity found by the user. Eventually, systems can report to designers on critical parts of their products. In both cases, intelligent systems can also be used to detect when hesitation appears to be a medical condition that impairs the user of interacting with the system in a standard manner, and therefore adapt their behaviour to allow, for instance, longer timers, simplified choices and assistance.

Applications may of course consider different cognitive or emotional factors that hesitation. Duric et al. [58] described how intelligent adaptive systems can respond to these factors: "*Imagine a computer interface that could predict and diagnose whether the user was fatigued, confused, frustrated, or momentarily distracted by gathering a variety of nonverbal information [...]. Further imagine that the interface could adapt itself - simplify, highlight, or tutor - to improve the humanâĂŞcomputer interaction (HCI) using these diagnoses and predictions. Nonverbal information facilitates a special type of communication where the goal is to probe the inner (cognitive and affective) states of the mind before any verbal communication has been contemplated and/or expressed.*". Machine learning to exploit the sensors' data is a potential to detect the user's cognitive and emotional behaviour that we encourage further researchers to implement in intelligent systems.

8

Conclusion

The topic of the correlation between gaze and hand during touch enabled devices interaction is cross-disciplinary. Essentially, research aims at explaining the correlation from a biomedical aspect: how the central nervous system organises the visual and motor commands to perform a manual reaching movement. In that sense, our work tried to give a general and basic description of the human behaviour observed in a specific context: the natural interaction with tablets. We described the correlation between gaze and hand at different stages of the hand activity commonly observed with users: taps, stationary hand events and motion in between the first two events. Whenever possible, we detailed the organisation of the gaze and hand during interaction according to the two aspects found in literature, temporally and spatially. Temporally, we confirmed that gaze leads the hand, it acquired the target area 159 pixels away by 338 ms before the hand had performed the tap. This result can serve as a benchmark for healthy users (mentally and physically) as our data collection relied on healthy users, with none to mild eye correction. Except for the limitations of eye tracking (cf. Section 3.1.4) that could discard some users, our study can be replicated with disabled users to measure their response against our baseline. However, our work did not allow to measure the *reaction* time between the gaze acquisition and the hand movement, because during the studies, there were no systematic stimulus nor resting position to measure this reaction time against. Spatially, our main finding was that the distance between gaze and hand (measured in the volume above the device's surface) is dependent of the location where the users looked at. The correlation was stronger horizontally, and it indicated that the

"mental" manual map of the interaction surface was a radial projection centred on the user's location, more distorted vertically, certainly to minimise efforts. We have shown that typing behaviour patterns are correlated with the user's typing skill (speed and accuracy of typing): good typist keep a larger vertical distance between their hand and their gaze, suggesting this is a very important factor to consider when a model of the correlation between gaze and hand needs to be made for usability by intelligent systems.

We have evaluated how the difficulty of the tasks impacted the correlation between gaze and hand during tablet interaction. Based on observations made on the levels of increasing difficulty in a Memory Game, we concluded that when the hand remained in a stationary phase, the distance between gaze and the projection of the hand on the tablet's screen increased with the difficulty, and that while the hand was in motion, the difficulty impacted mainly the horizontal alignment. Other factors also suggested that the increase of difficulty also deteriorated the correlation between gaze and hand. When the hand movements lasted longer, hand and gaze were less aligned. We suppose that longer movements were indicating less decisive actions. We have approximated the decisiveness of the participants by the success matching of the image pairs in the Memory Game, and found that when the stationary hand events were leading to an unsuccessful match, the distance between gaze and hand was larger.

The results we presented highlighted the individual differences found between participants. The correlation between gaze and taps clearly indicated that both the temporal and spatial dimensions of the correlation varied among users. These differences were expected: individuals moved their hand in different fashions probably related to experience (speed, reactivity), and certainly, in a natural context, connected with the different strategies and cognitive processes that influenced the way eyes and hands worked together.

Despite the noticeable individual differences, we have provided the core elements on which further work may build on, either to deepen the fundamental understanding of the eye/hand correlation when interacting with touch enabled devices, or to implement tools able to evaluate the behaviour adopted by touch enabled devices' users, and thus propose adequate and more intelligent responses from the systems towards enhanced interaction. Inspecting the machine learning to construct robust detection of emotional and cognitive behaviours is potential enhancement to the work we have done and that can be continued by further researchers.

In sum, we have devised two data collections to understand how gaze and hand are cor-

related during the interaction with touch-enabled devices, in a natural context (Internet related activities and games). Our population target was healthy subject with none to mild eye correction, and we focused on three hand events: taps, stationary hand events and hand in movement. We studied the impact of a few factors (typing, nature of the task, nature of the target). We have thought this research work as baseline to be exploited by intelligent systems to adapt themselves to the user's cognitive and emotional behaviours (intelligent Human-Computer Interaction [152] and adaptive systems principles [58, 122]) by proposing assistance or changes when such behaviours are detected, or provide tablet's application designers better understanding on their design pitfalls. We proposed the detection of indecision as an example. Our research was also thought to be implemented as public kiosks, hence the choice of a tablet and generic tasks. We suggest this work to be continued in the following ways: testing the applications, focusing on other hand events/gestures (i.e. zooms, drags), testing other devices configurations (i.e. change of orientation), focusing on other human factors (i.e. arousal, anxiety and engagement, and also by contrast, well-being) and comparing results with non standard subjects.

Bibliography

[1] Richard A. Abrams, David E. Meyer, and Sylvan Kornblum. Eye-hand coordination: oculomotor control in rapid aimed limb movements, 1990. [Cited on pages 20, 21, 24, and 112]

[2] Hyungil Ahn. *Modeling and analysis of affective influences on human experience, prediction, decision making, and behavior*. Phd thesis, MIT, 2010. [Cited on page 27]

[3] Arwa Alabdulkarim. Towards hand-gesture frustration detection in interactive systems. In *2014 3rd International Conference on User Science and Engineering (i-USEr)*, pages 153–157, sep 2014. [Cited on pages 4, 27, 84, and 121]

[4] Robert. Mc N. Alexander. A minimum energy cost hypothesis for human arm trajectories. *Biological Cybernetics*, 1997. [Cited on pages 105 and 111]

[5] Robert B. Allen. Mental models and user models. In *Handbook of Human-Computer Interaction*, pages 49–63. Elsevier, 1997. [Cited on page 110]

[6] Sherman R. Alpert, John Karat, Clare-Marie Karat, Carolyn Brodie, and John G. Vergo. User attitudes regarding a user-adaptive eCommerce web site. *User Modeling and User-Adapted Interaction*, 13(4):373–396, 2003. [Cited on page 35]

[7] Florian Alt, Alireza Sahami Shirazi, Albrecht Schmidt, and Julian Mennenöh. Increasing the user's attention on the web: using implicit interaction based on gaze behavior to tailor content. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, NordiCHI '12, pages 544–553, New York, NY, USA, 2012. ACM. [Cited on page 26]

[8] Ioannis Arapakis, Mounia Lalmas, and George Valkanas. Understanding within-content engagement through pattern analysis of mouse gestures. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1439–1448, New York, NY, USA, 2014. ACM. [Cited on page 27]

[9] Aristotle. *Parts of animals*. Loeb classical library. Harvard University Press ; Heinemann, Cambridge, MA : London, rev. edition, 1961. [Cited on page 15]

[10] Ernesto Arroyo, Ted Selker, and Willy Wei. Usability tool for analysis of web designs using mouse tracks. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 484–489, New York, NY, USA, 2006. ACM. [Cited on page 25]

[11] Richard Atterer and Albrecht Schmidt. Tracking the interaction of users with AJAX applications for usability testing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 1347–1350, New York, NY, USA, 2007. ACM. [Cited on pages 26 and 27]

[12] Richard Atterer, Monika Wnuk, and Albrecht Schmidt. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 203–212, New York, NY, USA, 2006. ACM. [Cited on pages 26 and 27]

[13] Anne Aula, Rehan M. Khan, and Zhiwei Guan. How does search behavior change as search becomes more difficult? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 35–44, New York, NY, USA, 2010. ACM. [Cited on page 25]

[14] David Benyon. Adaptive systems: A solution to usability problems. *User modeling and user-adapted interaction*, 3(1):65–87, mar 1993. [Cited on page 34]

[15] Hans-Joachim Bieg, Lewis L. Chuang, Roland W. Fleming, Harald Reiterer, and Heinrich H. Bülthoff. Eye and pointer coordination in search and selection tasks. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10, pages 89–92, New York, NY, USA, 2010. ACM. [Cited on pages 22, 23, 49, and 66]

[16] B. Biguer, Marc Jeannerod, and Claude Prablanc. The coordination of eye, head, and arm movements during reaching at a single visual target. *Experimental Brain Research*, 46(2):301–304, 1982. [Cited on pages 20, 24, and 112]

[17] Gordon Binsted, Romeo Chua, Werner F. Helsen, and Digby Elliott. Eye-hand coordination in goal-directed aiming. *Human Movement Science*, 20(4):563 – 585, 2001. [Cited on pages 20, 24, and 112]

[18] Otmar Bock. Contribution of retinal versus extraretinal signals towards visual localization in goal-directed movements. *Experimental Brain Research*, 64(3):476–482, nov 1986. [Cited on page 24]

[19] Thorsten Bohnenberger, Anthony Jameson, Antonio Krüger, and Andreas Butz. Location-aware shopping assistance: evaluation of a decision-theoretic approach. In Fabio Paternò, editor, *Human Computer Interaction with Mobile Devices*, pages 155–169, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. [Cited on page 35]

[20] Peter Brandl, Clifton Forlines, Daniel Wigdor, Michael Haller, and Chia Shen. Combining and measuring the benefits of bimanual pen and direct-touch interaction on horizontal interfaces. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '08, pages 154–161, New York, NY, USA, 2008. ACM. [Cited on page 2]

[21] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, sep 2002. [Cited on pages 25 and 26]

[22] Dermot Browne, Peter Totterdell, and Mike Norman, editors. *Adaptive User Interfaces*. Academic Press Ltd., London, UK, UK, 1990. [Cited on page 34]

[23] Roland Brünken, Jan L. Plass, and Detlev Leutner. Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1):53–61, mar 2003. [Cited on page 124]

[24] Paolo Burzacca and Fabio Paternò. Remote usability evaluation of mobile web applications. In Masaaki Kurosu, editor, *Human-Computer Interaction. Human-Centred Design Approaches, Methods, Tools, and Environments*, volume 8004 of *Lecture Notes in Computer Science*, pages 241–248. Springer Berlin Heidelberg, 2013. [Cited on page 26]

[25] Georg Buscher, Ralf Biedert, Daniel Heinesch, and Andreas Dengel. Eye tracking analysis of preferred reading regions on the screen. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 3307–3312, New York, NY, USA, 2010. ACM. [Cited on pages 26 and 27]

[26] Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 21–30, New York, NY, USA, 2009. ACM. [Cited on page 26]

[27] Georg Buscher, Susan T. Dumais, and Edward Cutrell. The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 42–49, New York, NY, USA, 2010. ACM. [Cited on pages 22 and 25]

[28] Georg Buscher, Ryen W. White, Susan Dumais, and Jeff Huang. Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings*

of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, pages 373–382, New York, NY, USA, 2012. ACM. [Cited on page 25]

[29] Cristian B. Calderon, Tom Verguts, and Wim Gevers. Losing the boundary: cognition biases action well after action selection. *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-GENERAL*, 144(4):737–743, 2015. [Cited on page 27]

[30] Les G. Carlton. Visual information: the control of aiming movements. *The Quarterly Journal of Experimental Psychology Section A*, 33(1):87–93, feb 1981. [Cited on pages 21 and 112]

[31] Craig S. Chapman, Jason P. Gallivan, Daniel K. Wood, Jennifer L. Milne, Jody C. Culham, and Melvyn A. Goodale. Reaching for the unknown: multiple target encoding and real-time decision-making in a rapid reach task. *Cognition*, 116(2):168–176, aug 2010. [Cited on page 21]

[32] Ishan Chatterjee, Robert Xiao, and Chris Harrison. Gaze+Gesture. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, ICMI '15, pages 131–138, New York, New York, USA, 2015. ACM Press. [Cited on page 29]

[33] Mon Chu Chen, John R. Anderson, and Myeong Ho Sohn. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, pages 281–282, New York, NY, USA, 2001. ACM. [Cited on pages 4, 22, 23, 27, and 49]

[34] Mon Chu Chen and Veraneka Lim. Eye gaze and mouse cursor relationship in a debugging task. In Constantine Stephanidis, editor, *HCI International 2013 - Posters' Extended Abstracts*, volume 373 of *Communications in Computer and Information Science*, pages 468–472. Springer Berlin Heidelberg, 2013. [Cited on page 23]

[35] Xiang 'Anthony' Chen, Julia Schwarz, Chris Harrison, Jennifer Mankoff, and Scott E. Hudson. Air+Touch: interweaving touch & in-air gestures. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 519–525, New York, NY, USA, 2014. ACM. [Cited on pages 31, 33, and 84]

[36] Pascal R. Chesnais, Matthew J. Mucklo, and Jonathan A. Sheena. The Fishwrap personalized news system. In *Proceedings of the Second International Workshop on Community Networking 'Integrated Multimedia Services to the Home'*, pages 275–282, 1995. [Cited on page 34]

[37] Victor Cheung, Jens Heydekorn, Stacey Scott, and Raimund Dachselt. Revisiting hovering: interaction guides for interactive surfaces. In *Proceedings of the 2012 ACM International*

Conference on Interactive Tabletops and Surfaces, ITS '12, pages 355–358, New York, NY, USA, 2012. ACM. [Cited on pages 32 and 84]

[38] Keith Cheverst, Nigel Davies, Keith Mitchell, Adrian Friday, and Christos Efstratiou. Developing a context-aware electronic tourist guide: some issues and experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 17–24, New York, NY, USA, 2000. ACM. [Cited on page 35]

[39] Sangwon Choi, Jaehyun Han, Sunjun Kim, Seongkook Heo, and Geehyuk Lee. ThickPad: a hover-tracking touchpad for a laptop. In *Proceedings of the 24th Annual ACM Symposium Adjunct on User Interface Software and Technology*, UIST '11 Adjunct, pages 15–16, New York, NY, USA, 2011. ACM. [Cited on pages 31 and 32]

[40] Kevin Christian, Bill Kules, Ben Shneiderman, and Adel Youssef. A comparison of voice controlled and mouse controlled web browsing. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies*, Assets '00, pages 72–79, New York, NY, USA, 2000. ACM. [Cited on page 27]

[41] Daniela Chudá and Peter Krátky. Usage of computer mouse characteristics for identification in web browsing. In *Proceedings of the 15th International Conference on Computer Systems and Technologies*, CompSysTech '14, pages 218–225, New York, NY, USA, 2014. ACM. [Cited on page 26]

[42] Mark Claypool, Phong Le, Makoto Wased, and David Brown. Implicit interest indicators. In *Proceedings of the 6th International Conference on Intelligent User Interfaces*, IUI '01, pages 33–40, New York, NY, USA, 2001. ACM. [Cited on pages 24 and 25]

[43] David R. Coghill, Sarah Seth, and Keith Matthews. A comprehensive assessment of memory, delay aversion, timing, inhibition, decision making and variability in attention deficit hyperactivity disorder: advancing beyond the three-pathway models. *Psychological medicine*, 44(9):1989–2001, jul 2014. [Cited on page 27]

[44] Han Collewijn and Ernst P. Tamminga. Human smooth and saccadic eye movements during voluntary pursuit of different target motions on different backgrounds. *The Journal of Physiology*, 351(1):217–250, 1984. [Cited on pages 111 and 116]

[45] Carlo Colombo, Alberto Del Bimbo, and Alessandro Valli. Visual capture and understanding of hand pointing actions in a 3-D environment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(4):677–686, 2003. [Cited on page 31]

[46] Arzu Çöltekin, Urska Demsar, Alzbeta Brychtova, and Jan Vandrol. Eye-hand coordination during visual search on geographic displays. In Peter Kiefer, Ioannis Giannopoulos, Martin

Raubal, and Antonio Krüger, editors, *Proceedings of the 2nd International Workshop on Eye Tracking for Spatial Research co-located with the 8th International Conference on Geographic Information Science, ET4S@GIScience 2014, Vienna, Austria, September 23, 2014.*, volume 1241 of *{CEUR} Workshop Proceedings*, pages 12–16. CEUR-WS.org, 2014. [Cited on pages 22 and 23]

[47] Lynne Cooke. Is the mouse a "poor man's eye tracker"? *Annual Conference-Society for Technical Communication*, 53:252, 2006. [Cited on pages 22, 23, and 26]

[48] Michael C. Corballis. *From hand to mouth: the origins of language.* Princeton University Press, Princeton ; Oxford, 2002. [Cited on page 15]

[49] François Courtemanche, Esma Aïmeur, Aude Dufresne, Mehdi Najjar, and Franck Mpondo. Activity recognition using eye-gaze movements and traditional interactions. *Interact. Comput.*, 23(3):202–213, may 2011. [Cited on page 27]

[50] Edward Cutrell and Zhiwei Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 407–416, New York, NY, USA, 2007. ACM. [Cited on page 25]

[51] Vagner Figuerêdo de Santana and Maria Cecília C. Baranauskas. Summarizing observational client-side data to reveal web usage patterns. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1219–1223, New York, NY, USA, 2010. ACM. [Cited on page 24]

[52] Fernando Diaz, Ryen White, Georg Buscher, and Dan Liebling. Robust models of mouse movement on dynamic web search results pages. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 1451–1460, New York, NY, USA, 2013. ACM. [Cited on page 25]

[53] Raymond Dodge and Thomas Sparks Cline. The angle velocity of eye movements. *Psychological Review*, 8(2):145–157, 1901. [Cited on page 13]

[54] Heiko Drewes. *Eye gaze tracking for human computer interaction.* PhD thesis, Faculty of Mathematics, Computer Science and Statistics, Ludwig-Maximilians-Universität München, March 2010. [Cited on page 14]

[55] Li Du, Yan Zhang, Chun-Chen Liu, Adrian Tang, Frank Hsiao, and Mau-Chung F. Chang. A 2.3-mW 11-cm range bootstrapped and correlated-double-sampling three-dimensional touch sensing circuit for mobile devices. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 64(1):96–100, jan 2017. [Cited on page 19]

[56] Andrew T. Duchowski. *Eye tracking methodology: theory and practice*. Springer-Verlag, Berlin, Heidelberg, 2007. [Cited on page 13]

[57] Susan T. Dumais, Georg Buscher, and Edward Cutrell. Individual differences in gaze patterns for web search. In *Proceedings of the Third Symposium on Information Interaction in Context*, IIiX '10, pages 185–194, New York, NY, USA, 2010. ACM. [Cited on pages 25 and 82]

[58] Zoran Duric, Wayne D. Gray, Ric Heishman, Fayin Li, Azriel Rosenfeld, Michael J. Schoelles, Christian Schunn, and Harry Wechsler. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proceedings of the IEEE*, 90(7):1272–1289, jul 2002. [Cited on pages 33, 36, 135, and 138]

[59] Claudia Ehmke and Stephanie Wilson. Identifying web usability problems from eye-tracking data. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 1*, BCS-HCI '07, pages 119–128, Swinton, UK, UK, 2007. British Computer Society. [Cited on page 26]

[60] Paul Ekman and Erika L. Rosenberg. *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford University Press, apr 2005. [Cited on page 131]

[61] Digby Elliott, Werner F. Helsen, and Romeo Chua. A century later: Woodworth's (1899) two-component model of goal-directed aiming. *Psychological bulletin*, 127(3):342–57, may 2001. [Cited on pages 20, 21, and 24]

[62] Burkhart Fischer and L. Rogal. Eye-hand-coordination in man: a reaction time study. *Biological Cybernetics*, 55(4):253–261, 1986. [Cited on pages 20 and 112]

[63] Gerhard Fischer. User modeling in human–computer interaction. *User Modeling and User-Adapted Interaction*, 11(1):65–86, mar 2001. [Cited on page 109]

[64] John D. Fisk and Melvyn A. Goodale. The organization of eye and limb movements during unrestricted reaching to targets in contralateral and ipsilateral visual space. *Experimental Brain Research*, 60(1), sep 1985. [Cited on pages 23 and 114]

[65] P. M. FITTS. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381–391, jun 1954. [Cited on page 124]

[66] Paul M. Fitts and Barbara K. Radford. Information capacity of discrete motor responses under different cognitive sets, 1966. [Cited on page 33]

[67] Tamar Flash. The control of hand equilibrium trajectories in multi-joint arm movements. *Biological Cybernetics*, 57(4-5):257–274, nov 1987. [Cited on page 15]

[68] Stephanie Foehrenbach, Werner A. König, Jens Gerken, and Harald Reiterer. Tactile feedback enhanced hand gesture interaction at large, high-resolution displays. *J. Vis. Lang. Comput.*, 20(5):341–351, 2009. [Cited on page 105]

[69] Nigel Ford, David Miller, and Nicola Moss. The role of individual differences in Internet searching: an empirical study. *J. Am. Soc. Inf. Sci. Technol.*, 52(12):1049–1066, oct 2001. [Cited on page 25]

[70] Tom Foulsham, Alan Kingstone, and Geoffrey Underwood. Turning the world around: patterns in saccade direction vary with picture orientation. *Vision Research*, 48(17):1777–1790, 2008. [Cited on pages 111, 116, and 128]

[71] Jason Friedman, Scott Brown, and Matthew Finkbeiner. Linking cognitive and reaching trajectories via intermittent movement control. *Journal of Mathematical Psychology*, 57(3):140–151, 2013. [Cited on page 121]

[72] Xin Fu. Towards a model of implicit feedback for web search. *J. Am. Soc. Inf. Sci. Technol.*, 61(1):30–49, jan 2010. [Cited on page 25]

[73] Krzysztof Z. Gajos, Mary Czerwinski, Mary Czerwinski, Desney S. Tan, and Daniel S. Weld. Exploring the design space for adaptive graphical user interfaces. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '06, pages 201–208, New York, NY, USA, 2006. ACM. [Cited on page 35]

[74] Iain D. Gilchrist and Monika Harvey. Evidence for a systematic component within scan paths in visual search. *Visual Cognition*, 14(4-8):704–715, 2006. [Cited on pages 111 and 116]

[75] Joseph H. Goldberg, Mark J. Stimson, Marion Lewenstein, Neil Scott, and Anna M. Wichansky. Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, ETRA '02, pages 51–58, New York, NY, USA, 2002. ACM. [Cited on page 27]

[76] Cleotilde Gonzalez. Does animation in user interfaces improve decision making? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, pages 27–34, New York, NY, USA, 1996. ACM. [Cited on page 28]

[77] Lee Griffiths and Zhongming Chen. Investigating the differences in web browsing behaviour of Chinese and European users using mouse tracking. In *Proceedings of the 2nd*

International Conference on Usability and Internationalization, UI-HCII'07, pages 502–512, Berlin, Heidelberg, 2007. Springer-Verlag. [Cited on pages 23 and 26]

[78] G. Mark Grimes and Joseph S. Valacich. Mind over mouse: the effect of cognitive load on mouse movement behavior. In *2015 International Conference on Information Systems: Exploring the Information Frontier, ICIS 2015*. Association for Information Systems, 2015. [Cited on page 125]

[79] Tovi Grossman, Daniel Wigdor, and Ravin Balakrishnan. Multi-finger gestural interaction with 3D volumetric displays. In *Proceedings of the 17th annual ACM symposium on User interface software and technology - UIST '04*, UIST '04, page 61, New York, New York, USA, 2004. ACM Press. [Cited on page 32]

[80] Peter Colin Groves and J. R. Napier. Encyclopaedia britannica, primate (hand and feet). `https://www.britannica.com/animal/primate-mammal/Hands-and-feet/`, 2018. [Online; last accessed Jan. 2019]. [Cited on page 15]

[81] Yves Guiard. On Fitts's and Hooke's laws: simple harmonic movement in upper-limb cyclical aiming. *Acta Psychologica*, 82(1-3):139–159, mar 1993. [Cited on page 23]

[82] Qi Guo and Eugene Agichtein. Towards predicting web searcher gaze position from mouse movements. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 3601–3606, New York, NY, USA, 2010. ACM. [Cited on pages 23, 25, 30, and 66]

[83] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. Towards estimating web search result relevance from touch interactions on mobile devices. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 1821–1826, New York, NY, USA, 2013. ACM. [Cited on page 25]

[84] Seungju Han and Joonah Park. A study on touch & hover based interaction for zooming. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, pages 2183–2188, New York, NY, USA, 2012. ACM. [Cited on pages 32 and 84]

[85] Jennifer Healey and Rosalind W. Picard. StartleCam: a cybernetic wearable camera. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*, pages 42–49, oct 1998. [Cited on page 132]

[86] Robert K. Heinssen, Carol R. Glass, and Luanne A. Knight. Assessing computer anxiety: development and validation of the computer anxiety rating scale. *Computers in Human Behavior*, 3(1):49–59, 1987. [Cited on page 132]

[87] Werner F. Helsen, Digby Elliott, Janet L. Starkes, and Kathryn L. Ricker. Temporal and spatial coupling of point of gaze and hand movements in aiming. *Journal of Motor Behavior*, 30(3):249–259, 1998. [Cited on page 21]

[88] Denise Y. P. Henriques and J. D. Crawford. Role of eye, head, and shoulder geometry in the planning of accurate arm movements. *Journal of Neurophysiology*, 87(4):1677–1685, apr 2002. [Cited on page 15]

[89] Ken Hinckley, Seongkook Heo, Michel Pahud, Christian Holz, Hrvoje Benko, Abigail Sellen, Richard Banks, Kenton O'Hara, Gavin Smyth, and William Buxton. Pre-touch sensing for mobile interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2869–2881, New York, NY, USA, 2016. ACM. [Cited on pages 19, 31, 32, and 112]

[90] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, 92(1-2):84–89, jun 2004. [Cited on page 131]

[91] Karen Ho and Hanley Weng. Favoured attributes of in-air gestures in the home environment. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, OzCHI '13, pages 171–174, New York, NY, USA, 2013. ACM. [Cited on page 105]

[92] Shuk Ying Ho and Kar Yan Tam. An empirical examination of the effects of web personalization at different stages of decision making. *International Journal of Human-Computer Interaction*, 19(1):95–112, sep 2005. [Cited on page 28]

[93] Corey Holland, Atenas Garza, Elena Kurtova, Jose Cruz, and Oleg Komogortsev. Usability evaluation of eye tracking on an unmodified common tablet. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 295–300, New York, NY, USA, 2013. ACM. [Cited on page 14]

[94] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. *Eye Tracking: a comprehensive guide to methods and measures*. OUP Oxford, 2011. [Cited on pages 12, 13, 41, and 93]

[95] Christoph Hölscher and Gerhard Strube. Web search behavior of Internet experts and newbies. *Comput. Netw.*, 33(1-6):337–346, jun 2000. [Cited on pages 25 and 26]

[96] Kristina Höök. Evaluating the utility and usability of an adaptive hypermedia system. *Knowledge-Based Systems*, 10(5):311–319, 1998. [Cited on page 34]

[97] Jeff Huang, Ryen White, and Georg Buscher. User see, user point: gaze and cursor alignment in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1341–1350, New York, NY, USA, 2012. ACM. [Cited on pages 22, 23, 25, 26, 49, and 66]

[98] George S. Hurst and William C. Colwell, Jr. Discriminating contact sensor, October 7 1975. US Patent 3,911,215. [Cited on page 18]

[99] Shamsi T. Iqbal and Brian P. Bailey. Using eye gaze patterns to identify user tasks. In *The Grace Hopper Celebration of Women in Computing 2004 (GHC 2004)*, 2004. [Cited on pages 26 and 36]

[100] Shamsi T. Iqbal, Xianjun Sam Zheng, and Brian P. Bailey. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, pages 1477–1480, New York, NY, USA, 2004. ACM. [Cited on page 36]

[101] Makio Ishihara and Yukio Ishihara. A reflex in eye-hand coordination for calibrating coordinates of a tabletop display. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, pages 297–300, New York, NY, USA, 2010. ACM. [Cited on page 23]

[102] Philip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*, pages 145–152, 2011. [Cited on page 110]

[103] Robert J. K. Jacob. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Trans. Inf. Syst.*, 9(2):152–169, apr 1991. [Cited on pages 1 and 29]

[104] Robert J.K. Jacob and Keith S. Karn. Eye tracking in human-computer interaction and usability research. In *The Mind's Eye*, pages 573–605. Elsevier, 2003. [Cited on page 1]

[105] Richard J. Jagacinski and John M. Flach. *Control theory for humans Quantitative approaches to modeling performance.* CRC Press, 2003. [Cited on page 15]

[106] Anthony Jameson. The Human-computer Interaction Handbook. chapter Adaptive interfaces and agents, pages 305–330. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003. [Cited on page 34]

[107] Bernard J. Jansen and Amanda Spink. *Web mining.* IGI Global, jan 2005. [Cited on page 68]

[108] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, 36(2):207–227, jan 2000. [Cited on page 25]

[109] Roger Johansson and Mikael Johansson. Look here, eye movements play a functional role in memory retrieval. *Psychological Science*, 25(1):236–242, 2014. [Cited on page 84]

[110] Eric A. Johnson. Touch display - a novel input/output device for computers. *Electronics Letters*, 1(8):219–220, 1965. [Cited on page 18]

[111] Kristiina Jokinen and Jens Allwood. *Hesitation in intercultural communication: some observations and analyses on interpreting shoulder shrugging*, pages 55–70. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. [Cited on page 131]

[112] Matt Jones, Gary Marsden, Norliza Mohd-Nasir, Kevin Boone, and George Buchanan. Improving web interaction on small displays. *Comput. Netw.*, 31(11-16):1129–1137, may 1999. [Cited on pages 24 and 26]

[113] Steven W. Keele and Michael I. Posner. Processing of visual feedback in rapid movements. *Journal of experimental psychology*, 77(1):155–8, may 1968. [Cited on pages 21, 23, and 24]

[114] Jaana Kekäläinen, Paavo Arvola, and Sanna Kumpulainen. Browsing patterns in retrieved documents. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 299–302, New York, NY, USA, 2014. ACM. [Cited on page 26]

[115] Melanie Kellar, Carolyn Watters, and Michael Shepherd. A field study characterizing web-based information-seeking tasks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7):999–1018, may 2007. [Cited on page 25]

[116] Werner A. König, Jens Gerken, Stefan Dierdorf, and Harald Reiterer. Adaptive pointing: implicit gain adaptation for absolute pointing devices. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, pages 4171–4176, New York, NY, USA, 2009. ACM. [Cited on page 16]

[117] Gerd Kortuem, Zary Segall, and Martin Bauer. Context-aware, adaptive wearable computers as remote interfaces to 'intelligent' environments. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*, pages 58–65, Oct 1998. [Cited on page 35]

[118] Hanna Maria Kaarina Koskinen, Jari Olavi Laarni, and Petri Mikael Honkamaa. Hands-on the process control: users preferences and associations on hand movements. In *CHI '08*

Extended Abstracts on Human Factors in Computing Systems, CHI EA '08, pages 3063–3068, New York, NY, USA, 2008. ACM. [Cited on pages 105 and 116]

[119] Kai Kunze, Shoya Ishimaru, Yuzuko Utsumi, and Koichi Kise. My reading life: towards utilizing eyetracking on unmodified tablets and phones. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, UbiComp '13 Adjunct, pages 283–286, New York, NY, USA, 2013. ACM. [Cited on page 14]

[120] Dmitry Lagun, Mikhail Ageev, Qi Guo, and Eugene Agichtein. Discovering common motifs in cursor movement data for improving web search. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 183–192, New York, NY, USA, 2014. ACM. [Cited on pages 24 and 25]

[121] Michael F. Land. Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 25(3):296–324, may 2006. [Cited on pages 1 and 23]

[122] Pat Langley. User modeling in adaptive interfaces. In *UM99 User Modeling*, pages 357–370, 1999. [Cited on pages 110, 111, and 138]

[123] Daniel J. Liebling and Susan T. Dumais. Gaze and mouse coordination in everyday work. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 1141–1150, New York, NY, USA, 2014. ACM. [Cited on pages 1, 22, 23, 49, 66, 67, and 112]

[124] Irene Lopatovska and Ioannis Arapakis. Theories, methods and current research on emotions in library and information science, information retrieval and humanâĂŞcomputer interaction. *Information Processing & Management*, 47(4):575–592, 2011. [Cited on page 111]

[125] Yannick Lufimpu-Luviya, Djamel Merad, Sebastien Paris, Véronique Drai-Zerbib, Thierry Baccino, and Bernard Fertil. A regression-based method for the prediction of the indecisiveness degree through eye movement patterns. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*, ETSA '13, pages 32–38, New York, NY, USA, 2013. ACM. [Cited on pages 27 and 110]

[126] I. Scott MacKenzie. Human-computer interaction. chapter Movement T, pages 483–492. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995. [Cited on page 125]

[127] Ian Scott MacKenzie. Fitts' law as a research and design tool in Human-Computer Interaction. *HumanâĂŞComputer Interaction*, 7(1):91–139, 1992. [Cited on page 125]

[128] Nicolai Marquardt, Ricardo Jota, Saul Greenberg, and Joaquim A. Jorge. The continuous interaction space: interaction techniques unifying touch and gesture on and above a digital surface. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part III*, INTERACT'11, pages 461–476, Berlin, Heidelberg, 2011. Springer-Verlag. [Cited on pages 31 and 32]

[129] Joanna McGrenere, Ronald M. Baecker, and Kellogg S. Booth. An evaluation of a multiple interface design solution for bloated software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 164–170, New York, NY, USA, 2002. ACM. [Cited on page 35]

[130] David E. Meyer, J. E. Keith Smith, Sylvan Kornblum, Richard A. Abrams, and Charles E. Wright. Speed-accuracy tradeoffs in aimed movements: Toward a theory of rapid voluntary action. In *Attention and performance 13: Motor representation and control*, pages 173–226. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1990. [Cited on page 33]

[131] Dragana Micic, Howard Ehrlichman, and Rebecca Chen. Why do we move our eyes while trying to remember? The relationship between non-visual gaze patterns and memory. *Brain and Cognition*, 74(3):210–224, 2010. [Cited on page 106]

[132] AJung Moon, Chris A. C. Parker, Elizabeth A. Croft, and H. F. Machiel Van der Loos. Did you see it hesitate? - Empirically grounded design of hesitation trajectories for collaborative robots. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1994–1999. IEEE, sep 2011. [Cited on page 33]

[133] Pietro Morasso. Spatial control of arm movements. *Experimental Brain Research*, 42(2):223–227, apr 1981. [Cited on page 15]

[134] Steven Morrison and Justin Keogh. Changes in the dynamics of tremor during goal-directed pointing. *Human Movement Science*, 20(4):675–693, 2001. [Cited on page 16]

[135] Steven Morrison and Karl M. Newell. Postural and resting tremor in the upper limb. *Clinical Neurophysiology*, 111(4):651–663, 2000. [Cited on page 16]

[136] Xiangwei Mu, Yan Chen, Jian Yang, and Jingjing Jiang. An improved similarity algorithm based on hesitation degree for user-based collaborative filtering. In *Proceedings of the 5th International Conference on Advances in Computation and Intelligence*, ISICA'10, pages 261–271, Berlin, Heidelberg, 2010. Springer-Verlag. [Cited on page 33]

[137] Christian Mühl, Brendan Allison, Anton Nijholt, and Guillaume Chanel. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 1(2):66–84, 2014. [Cited on page 132]

[138] Hendrik Müller, Jennifer Gove, and John Webb. Understanding tablet use: a multi-method exploration. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '12, pages 1–10, New York, NY, USA, 2012. ACM. [Cited on page 133]

[139] Christian Müller-Tomfelde. Dwell-Based Pointing in Applications of Human Computer Interaction. In Cécilia Baranauskas, Philippe Palanque, Julio Abascal, and Simone Diniz Junqueira Barbosa, editors, *Human-Computer Interaction – INTERACT 2007*, pages 560–573, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. [Cited on page 31]

[140] Noboru Nakamichi, Kazuyuki Shima, Makoto Sakai, and Ken-ichi Matsumoto. Detecting low usability web pages using quantitative data of users' behavior. In *Proceedings of the 28th International Conference on Software Engineering*, ICSE '06, pages 569–576, New York, NY, USA, 2006. ACM. [Cited on page 26]

[141] Vidhya Navalpakkam and Elizabeth Churchill. Mouse tracking: measuring and predicting users' experience of web-based content. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2963–2972, New York, NY, USA, 2012. ACM. [Cited on pages 23 and 27]

[142] Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 953–964, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee. [Cited on pages 23, 25, and 30]

[143] Sebastiaan F. W. Neggers and Harold Bekkering. Coordinated control of eye and hand movements in dynamic reaching. *Human Movement Science*, 21(3):37–64, sep 2002. [Cited on pages 23 and 24]

[144] Jakob Nielsen. Horizontal attention leans left. `https://www.nngroup.com/articles/horizontal-attention-leans-left/`, 2010. [Online; last accessed Jan. 2019]. [Cited on pages 71 and 82]

[145] Heather L. O'Brien and Elaine G. Toms. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955, 2008. [Cited on page 131]

[146] Hayato Ohmura, Teruaki Kitasuka, and Masayoshi Aritsugi. A web browsing behavior recording system. In Andreas König, Andreas Dengel, Knut Hinkelmann, Koichi Kise,

Robert J. Howlett, and Lakhmi C. Jain, editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6884 of *Lecture Notes in Computer Science*, pages 53–62. Springer Berlin Heidelberg, 2011. [Cited on pages 25 and 27]

[147] Takeo Onishi and Takahiro Shiroshima. Predicting touch operations by using hover information in smartphones for data prefetching. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, jul 2016. [Cited on pages 32 and 112]

[148] Anna Ostberg and Nada Matic. Hover cursor: improving touchscreen acquisition of small targets with hover-enabled pre-selection. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1723–1728, New York, NY, USA, 2015. ACM. [Cited on pages 32, 33, and 84]

[149] Sharon Oviatt. Human-centered design meets Cognitive Load Theory: designing interfaces that help people think. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM '06, pages 871–880, New York, NY, USA, 2006. ACM. [Cited on page 34]

[150] Fred Paas, Juhani E. Tuovinen, Huib Tabbers, and Pascal W. M. Van Gerven. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1):63–71, mar 2003. [Cited on page 124]

[151] Bing Pan, Helene A. Hembrooke, Geri K. Gay, Laura A. Granka, Matthew K. Feusner, and Jill K. Newman. The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, ETRA '04, pages 147–154, New York, NY, USA, 2004. ACM. [Cited on page 27]

[152] Maja Pantic, Anton Nijholt, Alex Pentland, and Thomas S. Huanag. Human-centred intelligent human-computer interaction (hci^2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008. [Cited on pages 3, 33, 34, 135, and 138]

[153] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA '18, pages 21:1–21:10, New York, NY, USA, 2018. ACM. [Cited on page 14]

[154] Andrea L. Patalano, Barbara J. Juhasz, and Joanna Dicke. The relationship between indecisiveness and eye movement patterns in a decision making informational search task.

Journal of Behavioral Decision Making, 23(4):353–368, 2010. [Cited on pages 27, 85, 110, and 111]

[155] Jon Peddie. *The history of visual magic in computers*. Springer London, London, 2013. [Cited on page 18]

[156] Jeff Pelz, Mary Hayhoe, and Russ Loeber. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, 139(3):266–277, aug 2001. [Cited on pages 1 and 112]

[157] Lucas Pereira. A mouse (h)over a hotspot survey: an exploration of patterns of hesitation through cursor movement metrics. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages LBW1522:1—-LBW1522:6, New York, NY, USA, 2019. ACM. [Cited on page 111]

[158] 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, New York, NY, USA, 2015. ACM. [Cited on page 1]

[159] Ken Pfeuffer, Jason Alexander, and Hans Gellersen. GazeArchers: playing with individual and shared attention in a two-player look & shoot Tabletop Game. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia*, MUM '16, pages 213–216, New York, NY, USA, 2016. ACM. [Cited on page 1]

[160] Ken Pfeuffer, Jason Alexander, and Hans Gellersen. Partially-indirect bimanual input with gaze, pen, and touch for pan, zoom, and ink interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2845–2856, New York, NY, USA, 2016. ACM. [Cited on page 1]

[161] Ken Pfeuffer and Hans Gellersen. Gaze and touch interaction on tablets. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pages 301–311, New York, NY, USA, 2016. ACM. [Cited on pages 1 and 30]

[162] Ken Pfeuffer, Ken Hinckley, Michel Pahud, and Bill Buxton. Thumb + Pen interaction on tablets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3254–3266, New York, NY, USA, 2017. ACM. [Cited on page 1]

[163] R W Picard and J Healey. Affective wearables. In *Digest of Papers. First International Symposium on Wearable Computers*, pages 90–97, oct 1997. [Cited on page 132]

[164] Siddharth S. Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54, jan 2015. [Cited on page 17]

[165] Bryan Reimer. Impact of cognitive task complexity on drivers' visual tunneling. *Transportation Research Record*, 2138(1):13–19, 2009. [Cited on page 119]

[166] David A. Robinson. A method of measuring eye movemnent using a scieral search coil in a magnetic field. *IEEE Transactions on Bio-medical Electronics*, 10(4):137–145, oct 1963. [Cited on page 13]

[167] Kerry Rodden, Xin Fu, Anne Aula, and Ian Spiro. Eye-mouse coordination patterns on web search results pages. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 2997–3002, New York, NY, USA, 2008. ACM. [Cited on pages 22, 23, and 25]

[168] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 13–19, New York, NY, USA, 2004. ACM. [Cited on page 25]

[169] Uta Sailer, Thomas Eggert, Jochen Ditterich, and Andreas Straube. Spatial and temporal aspects of eye-hand coordination across different tasks. *Experimental Brain Research*, 134(2):163–173, sep 2000. [Cited on page 1]

[170] 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, New York, NY, USA, 2000. ACM. [Cited on pages 85 and 89]

[171] E. Schott. Über die Registrierung des Nystagmus und anderer Augenbewegungen vermittels des Seitengalvenometers. *Deutsches Archiv fur Klinische Medizin*, 40:79–90, 1922. [Cited on page 14]

[172] Sujan Shrestha. Mobile web browsing: usability study. In *Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology*, Mobility '07, pages 187–194, New York, NY, USA, 2007. ACM. [Cited on page 26]

[173] Dana Slambekova, Reynold Bailey, and Joe Geigel. Gaze and gesture based object manipulation in virtual worlds. In *Proceedings of the 18th ACM symposium on Virtual reality software and technology - VRST '12*, VRST '12, page 203, New York, New York, USA, 2012. ACM Press. [Cited on page 29]

[174] Barton A. Smith, Janet Ho, Wendy Ark, and Shumin Zhai. Hand eye coordination patterns in target selection. In *Proceedings of the symposium on Eye tracking research & applications - ETRA '00*, ETRA '00, pages 117–122, New York, New York, USA, 2000. ACM Press. [Cited on pages 22 and 66]

[175] Joo-Hyun Song and Ken Nakayama. Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, 13(8):360–366, 2009. [Cited on pages 21 and 121]

[176] William A. Sparrow and Karl M. Newell. Metabolic energy expenditure and the regulation of movement economy. *Psychonomic Bulletin & Review*, 5(2):173–196, Jun 1998. [Cited on page 21]

[177] Cheri Speier, Joseph S. Valacich, and Iris Vessey. The influence of task interruption on individual decision making: an information overload perspective. *Decision Sciences*, 30(2):337–360, mar 1999. [Cited on page 27]

[178] Sophie Stellmach and Raimund Dachselt. Designing gaze-based user interfaces for steering in virtual environments. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, ETRA '12, page 131, New York, New York, USA, 2012. ACM Press. [Cited on page 1]

[179] 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*, ETRA '12, page 357, New York, New York, USA, 2012. ACM Press. [Cited on page 1]

[180] Sophie Stellmach and Raimund Dachselt. Still looking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, CHI '13, page 285, New York, New York, USA, 2013. ACM Press. [Cited on page 1]

[181] Steven Strachan and Roderick Murray-Smith. Muscle tremor as an input mechanism. *UIST*, 2004. [Cited on page 16]

[182] Derek F. Stubbs. What the eye tells the hand. *Journal of Motor Behavior*, 8(1):43–58, mar 1976. [Cited on pages 15 and 21]

[183] Hideaki Takahira, Kei Kikuchi, and Mitsuho Yamada. A system for measuring gaze movement and hand movement simultaneously for hand-held devices. *IEICE Transactions on Communications*, E98.B(1):51–61, 2015. [Cited on page 30]

[184] Desney Tan and Anton Nijholt. *Brain-Computer Interfaces and Human-Computer Interaction*, pages 3–19. Springer London, London, 2010. [Cited on page 36]

[185] Gek Woo Tan and Kwok Kee Wei. An empirical study of web browsing behaviour: towards an effective website design. *Electronic Commerce Research and Applications*, 5(4):261–271, 2006. [Cited on page 26]

[186] Sivanagaraja Tatinati, Yubo Wang, and Kalyana C. Veluvolu. Modeling of physiological tremor with quaternion variant of extreme learning machines. In *Proceedings of the 2nd International Conference on Communication and Information Processing*, ICCIP '16, pages 255–258, New York, NY, USA, 2016. ACM. [Cited on page 16]

[187] Benjamin W. Tatler, Mary M. Hayhoe, Michael F. Land, and Dana H. Ballard. Eye guidance in natural vision: reinterpreting salience. *Journal of vision*, 11(5):5, may 2011. [Cited on page 81]

[188] Tobias Tempel and Christian Frings. How motor practice shapes memory: retrieval but not extra study can cause forgetting. *Memory (Hove, England)*, 24(7):903–915, aug 2016. [Cited on page 84]

[189] Alan J. Thurston. Giovanni Borelli and the study of human movement: an historical review. *ANZ Journal of Surgery*, 69(4):276–288, apr 1999. [Cited on page 16]

[190] Jayson Turner. Cross-device eye-based interaction. In *Proceedings of the Adjunct Publication of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13 Adjunct, pages 37–40, New York, NY, USA, 2013. ACM. [Cited on page 1]

[191] Jayson Turner, Andreas Bulling, Jason Alexander, and Hans Gellersen. Cross-device gaze-supported point-to-point content transfer. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '14, pages 19–26, New York, NY, USA, 2014. ACM. [Cited on pages 1 and 30]

[192] Jayson Turner, Andreas Bulling, and Hans Gellersen. Combining gaze with manual interaction to extend physical reach. In *Proceedings of the 1st International Workshop on Pervasive Eye Tracking & Mobile Eye-based Interaction*, PETMEI '11, pages 33–36, New York, NY, USA, 2011. ACM. [Cited on pages 1 and 30]

[193] Jayson Turner, Shamsi Iqbal, and Susan Dumais. Understanding gaze and scrolling strategies in text consumption tasks. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, UbiComp/ISWC'15 Adjunct, pages 829–838, New York, NY, USA, 2015. ACM. [Cited on page 124]

[194] Yoji Uno, Mitsuo Kawato, and R Suzuki. Formation and control of optimal trajectory in human multijoint arm movement. *Biological Cybernetics*, 61(2):89–101, 1989. [Cited on page 15]

[195] Krist Vaesen. The cognitive bases of human tool use. *Behavioral and Brain Sciences*, 35(4):203–218, aug 2012. [Cited on page 2]

[196] Cornelis J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979. [Cited on page 95]

[197] Eduardo Velloso, Jayson Turner, Jason Alexander, Andreas Bulling, and Hans Gellersen. An empirical investigation of gaze selection in mid-air gestural 3D manipulation. In Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler, editors, *Human-Computer Interaction – INTERACT 2015*, pages 315–330, Cham, 2015. Springer International Publishing. [Cited on page 29]

[198] Jean-Louis Vercher, Giovanni Magenes, Claude Prablanc, and Gabriel M. Gauthier. Eye-head-hand coordination in pointing at visual targets: spatial and temporal analysis. *Experimental Brain Research*, 99(3):507–523, Jan 1994. [Cited on pages 15, 20, 24, and 112]

[199] Eric D. Vidoni, Jason S. McCarley, Jodi D. Edwards, and Lara A. Boyd. Manual and oculomotor performance develop contemporaneously but independently during continuous tracking. *Experimental Brain Research*, 195(4):611–620, jun 2009. [Cited on page 1]

[200] Tomaž Vodlan and Andrej Košir. Methodology for transformation of behavioural cues into social signals in human-computer interaction. In Francisco V. Cipolla-Ficarra, editor, *Handbook of Research on Interactive Information Quality in Expanding Social Network Communications*, pages 104–118. IGI Global, Hershey, PA, USA, 2015. [Cited on page 33]

[201] Tomaž Vodlan, Marko Tkalčič, and Andrej Košir. The impact of hesitation, a social signal, on a user's quality of experience in multimedia content retrieval. *Multimedia Tools Appl.*, 74(17):6871–6896, sep 2015. [Cited on pages 33 and 121]

[202] Daniel Vogel and Ravin Balakrishnan. Distant freehand pointing and clicking on very large, high resolution displays. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, UIST '05, pages 33–42, New York, NY, USA, 2005. ACM. [Cited on page 93]

[203] Chat Wacharamanotham, Kashyap Todi, Marty Pye, and Jan Borchers. Understanding finger input above desktop devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1083–1092, New York, NY, USA, 2014. ACM. [Cited on pages 31, 32, and 84]

[204] Nicholas J. Wade. *Moving tablet of the eye : the origins of modern eye movement research.* Oxford University Press, Oxford, 2005. [Cited on page 13]

[205] Brian A. Wandell. *Foundations of vision.* Sinauer Associates, Sunderland, Mass., 1995. [Cited on pages 11, 12, and 116]

[206] Peiling Wang, William B. Hawk, and Carol Tenopir. Users' interaction with world wide web resources: an exploratory study using a holistic approach. *Inf. Process. Manage.*, 36(2):229–251, jan 2000. [Cited on page 26]

[207] Jamie A. Ward, Paul Lukowicz, and Hans W. Gellersen. Performance metrics for activity recognition. *ACM Trans. Intell. Syst. Technol.*, 2(1):6:1—-6:23, 2011. [Cited on page 94]

[208] Pierre Weill-Tessier and Hans Gellersen. Correlation between gaze and hovers during decision-making interaction. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications - ETRA '18*, ETRA '18, pages 1–5, New York, New York, USA, 2018. ACM Press. [Not cited]

[209] Pierre Weill-Tessier and Hans Gellersen. Touch input and gaze correlation on tablets. In Ireneusz Czarnowski, Robert J. Howlett, and Lakhmi C. Jain, editors, *Intelligent Decision Technologies 2017: Proceedings of the 9th KES International Conference on Intelligent Decision Technologies (KES-IDT 2017) – Part II*, pages 287–296. Springer International Publishing, Cham, 2018. [Not cited]

[210] Pierre Weill-Tessier, Jayson Turner, and Hans Gellersen. How do you look at what you touch? In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications - ETRA '16*, ETRA '16, pages 329–330, New York, New York, USA, 2016. ACM Press. [Cited on page 84]

[211] Ryen W. White and Dan Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 255–262, New York, NY, USA, 2007. ACM. [Cited on pages 24 and 25]

[212] Jeffery D. Wilfong. Computer anxiety and anger: the impact of computer use, computer experience, and self-efficacy beliefs. *Computers in Human Behavior*, 22(6):1001–1011, 2006. [Cited on page 132]

[213] Andrew D. Wilson and Hrvoje Benko. Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 273–282, New York, NY, USA, 2010. ACM. [Cited on pages 31 and 32]

[214] Robert S. Woodworth. Accuracy of voluntary movement. *The Psychological Review: Monograph Supplements*, 3(3):i–114, 1899. [Cited on pages 20 and 23]

[215] Haijun Xia, Ricardo Jota, Benjamin McCanny, Zhe Yu, Clifton Forlines, Karan Singh, and Daniel Wigdor. Zero-latency tapping: using hover information to predict touch locations and eliminate touchdown latency. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 205–214, New York, NY, USA, 2014. ACM. [Cited on pages 32 and 112]

[216] Ling-lin Xia, Bing Zou, Hao Liu, Hai Su, and Huang Qianghui. A new method for evaluating postural hand tremor based on CMOS camera. *Optik - International Journal for Light and Electron Optics*, 126(5):507–512, 2015. [Cited on pages 16 and 93]

[217] Pingmei Xu, Yusuke Sugano, and Andreas Bulling. Spatio-temporal modeling and prediction of visual attention in graphical user interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, CHI '16, pages 3299–3310, New York, New York, USA, 2016. ACM Press. [Cited on page 66]

[218] Che-Chang Yang and Yeh-Liang Hsu. A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors (Basel, Switzerland)*, 10(8):7772–7788, 2010. [Cited on page 131]

[219] Xing-Dong Yang, Tovi Grossman, Pourang Irani, and George Fitzmaurice. TouchCuts and TouchZoom: enhanced target selection for touch displays using finger proximity sensing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2585–2594, New York, NY, USA, 2011. ACM. [Cited on page 112]

[220] ByungIn Yoo, Jae-Joon Han, Changkyu Choi, Kwonju Yi, Sungjoo Suh, Dusik Park, and Changyeong Kim. 3D user interface combining gaze and hand gestures for large-scale display. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, CHI EA '10, page 3709, New York, New York, USA, 2010. ACM Press. [Cited on page 29]

[221] Wayne W. Zachary and Joan M. Ryder. Chapter 52 - Decision support systems: integrating decision aiding and decision training. *Handbook of Human-Computer Interaction*, pages 1235–1258, 1997. [Cited on page 27]

[222] Shumin Zhai, Carlos Morimoto, and Steven Ihde. Manual and gaze input cascaded (MAGIC) pointing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 246–253, New York, NY, USA, 1999. ACM. [Cited on pages 1 and 29]

[223] Yanxia Zhang, Andreas Bulling, and Hans Gellersen. Sideways: a gaze interface for spontaneous interaction with situated displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 851–860, New York, NY, USA, 2013. ACM. [Cited on pages 1 and 14]

[224] Yanxia Zhang, Sophie Stellmach, Abigail Sellen, and Andrew Blake. The costs and benefits of combining gaze and hand gestures for remote interaction. In Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler, editors, *Human-Computer Interaction – INTERACT 2015*, pages 570–577, Cham, 2015. Springer International Publishing. [Cited on page 30]

Appendix

A Gaze and Tap Correlation Study Material

Table A.1: Questions of the Search Task.

1 (I)	What are some side-effects of Ibuprofen?
2 (N)	Find the special offers page for Southwest Airlines.
3 (N)	Find the homepage of the "Pinewood" software company.
4 (N)	Find the homepage of the Football World Cup 2006.
5 (N)	Find the homepage of the School of Computing and Communications of Lancaster University.
6 (I)	What is the size of a modern implantable pacemaker of today?
7 (I)	I was watching the movie "Stand by Me" the other day. I know it is based on a Stephen King story with a different name. What is the name of the story?
8 (I)	Find the names of Julia Roberts' children.
9 (N)	Find the official website of Tesla Motors - a startup that builds powerful electronic cars.
10 (I)	A technician cuts his finger badly in the Biology Department. What are the legal implications of this for the university? Find relevant information on the Web.

(I) Informational, (N) Navigational

Table A.2: Suggestion of Mock-up Personal Data for the Shopping Task.

Name	Ashley Simpson
Age	25
Address	InfoLab21 South Drive Lancaster LA1 4WA
Email	a.simpson@email.com
Phone	079 12345678

Table A.3: Source and Target Articles (2 rounds) for the Game Task.

	Source	*Target*
First round	**Pottery** -	**Martin Van Buren** Martin Van Buren (1782-1862) was the eighth President of the United States (1837-1841). Before his presidency, he was the eighth Vice President (1833âĂŞ1837) and the tenth secretary of state (1829-1831), both under Andrew Jackson. Van Buren was a key organizer of the Democratic Party, a dominant figure in the Second Party System, and the first president not of British or Irish descent—his family was Dutch. His administration was largely characterized by the economic hardship of his time, the Panic of 1837. Van Buren was the last Vice President to be elected directly to the presidency until George H. W. Bush in 1988.
Second round	**Eureka Tower** Eureka Tower is a skyscraper located in the Southbank precinct of Melbourne, Victoria, Australia. The project was designed by Melbourne architectural firm Fender Katsalidis Architects and was built by Grocon (Grollo Australia). The developer of the tower was Eureka Tower Pty Ltd, a joint venture consisting of Daniel Grollo (Grocon), investor Tab Fried and one of the Tower's architects Nonda Katsalidis. As of December 2013 it is the 14th tallest residential building in the world. It is currently the 98th tallest building in the world.	**Abbot's booby** The Abbott's booby (Papasula abbotti) is an endangered seabird. Abbott's booby breeds only in a few spots on the Australian territory of Christmas Island in the eastern Indian Ocean, although it formerly had a much wider range. It has white plumage with black markings, and is adapted for long-distance flight.

Questionnaire

Participant ID: _____

About you...

Gender: ☐ Male ☐ Female

Professional background: _____

Age: _____

Dominant hand: ☐ Left ☐ Right

Eye sight correction: ☐ None ☐ Glasses ☐ Contact Lenses

English language comprehension: [poor] 1 2 3 4 5 [excellent]

About your experience...

How would you define your experience with tablets? [inexperienced] 1 2 3 4 5 [expert]

How would you define your experience with enabled multitouch devices (such as iPhone, Android ...)? [inexperienced] 1 2 3 4 5 [expert]

About this experiment...

How would you evaluate the usability of the emulated browser (the application) used for this experiment? [bad] 1 2 3 4 5 [good]

If possible, can you detail the reason for your answer's choice at the previous question?

_____

_____

_____

_____

_____

_____

_____

How would you evaluate the usability of the overall layout used for this experiment? [bad] 1 2 3 4 5 [good]

If possible, can you detail the reason for your answer's choice at the previous question?

_____

_____

_____

_____

_____

_____

_____

Can you state any other comments, critics (good or bad) or suggestions?

_____

_____

_____

_____

_____

_____

_____

THANK YOU VERY MUCH FOR YOUR PARTICIPATION

Figure A.1: Questionnaire submitted at the end of the study.

```
private const string GET_ACTIVE_ELEMENT_POSITION_SCRIPT =
    "function extraScriptGetElementX() { return document.activeElement.getBoundingClientRect().left; }"+
    "function extraScriptGetElementY() { return document.activeElement.getBoundingClientRect().top; }";
private const string GET_CURSOR_SCREEN_POSITION_SCRIPT =
    "var _cursorScreenX=-1; var _cursorScreenY=-1;" +
    "document.onclick=setCursorPos;" +
    "function setCursorPos(obj){ _cursorScreenX=obj.screenX; _cursorScreenY=obj.screenY; }" +
    "function extraScriptGetCursorX() { return _cursorScreenX; }" +
    "function extraScriptGetCursorY() { return _cursorScreenY; }";

//...

protected void webBrowser_DocumentCompleted(object sender, WebBrowserDocumentCompletedEventArgs e)
{
    //....
    //script for retrieving an element relative to the viewport coordinates
    HtmlElement scriptEl = webBrowser.Document.CreateElement("script");
    ((IHTMLScriptElement)scriptEl.DomElement).text = GET_ACTIVE_ELEMENT_POSITION_SCRIPT + GET_CURSOR_SCREEN_POSITION_SCRIPT;
    webBrowser.Document.GetElementsByTagName("HEAD")[0].AppendChild(scriptEl);

    if (_firstWebBrowserLoad)
    {
        webBrowser.Document.Click += OnWebBrowserDoClick;
        _firstWebBrowserLoad = false;
    }
    //...
}

private void OnWebBrowserDoClick(object sender, HtmlElementEventArgs e)
{
    _dtEvent = DateTime.Now;
    HtmlElement currentElmt = webBrowser.Document.ActiveElement;
    if (currentElmt != null)
    {
        File.AppendAllText(HTML_LOG_FILE, string.Format("{0}\t{1}\t{2}\t{3}\t{4}\t{5}\t{6}\t{7}\t{8}\t{9}\n",
            _dtEvent.ToString("yyyy/MM/dd HH:mm:ss.ffff"),
            _dtEvent.Ticks,
            e.EventType,
            currentElmt.TagName,
            currentElmt.ScrollRectangle.Width,
            currentElmt.ScrollRectangle.Height,
            webBrowser.Document.InvokeScript("extraScriptGetElementX"),
            webBrowser.Document.InvokeScript("extraScriptGetElementY"),
            //these are the ABSOLUTE coordinates to the screen
            e.ClientMousePosition.X + webBrowser.Location.X,
            e.ClientMousePosition.Y + webBrowser.Location.Y));
        //...
    }
}
```

Figure A.2: Snippet of the JavaScript code injection on the webpages on the emulated browser.

Figure A.3: Flyer of the data collection participation (related to Chapter 4).

B Gaze and Hovers Correlation Memory Game Details

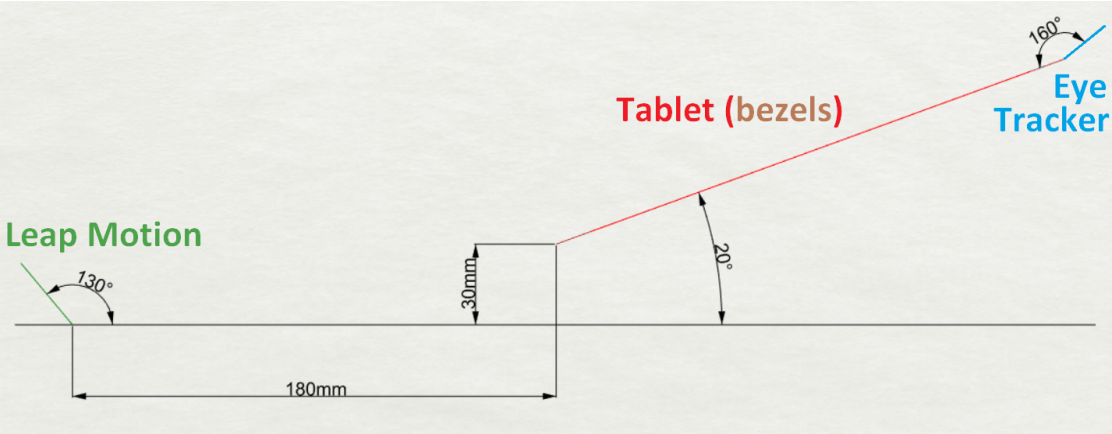


Figure B.1: Schematic disposition of the apparatus elements. (sideways view)

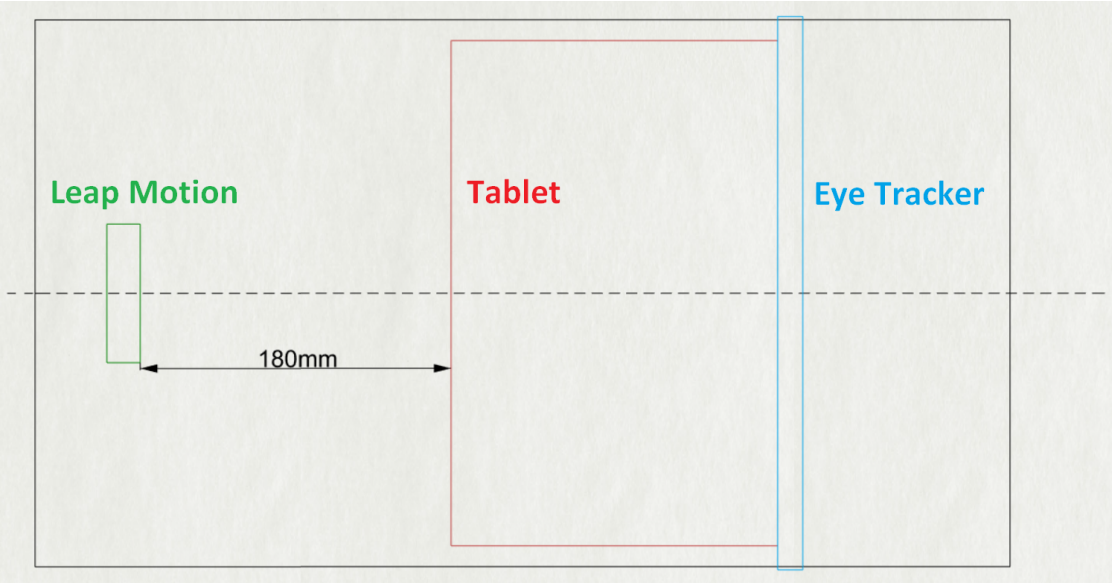


Figure B.2: Schematic disposition of the apparatus elements. (top view, alignment is suggested by the dotted line)

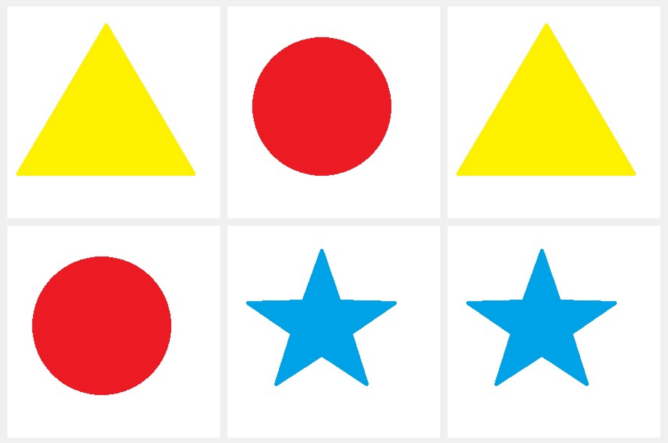


Figure B.3: Demonstration Version of the Memory Game.

Figure B.4: Level 1 of the Memory Game.

Figure B.5: Level 2 of the Memory Game.

Figure B.6: Level 3 of the Memory Game.

Figure B.7: Flyer of the data collection participation (related to Chapters 5 and 6).

C Gaze and Hovers Correlation Other Results

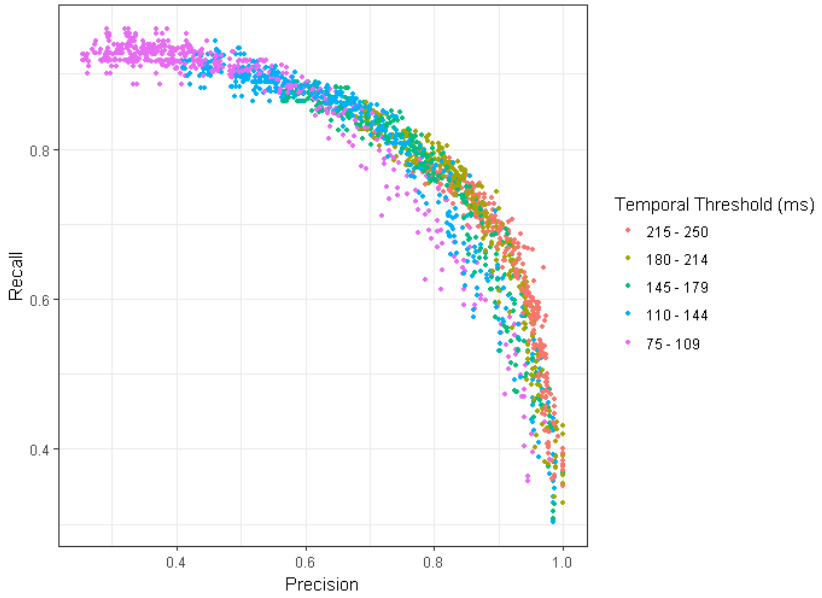


Figure C.1: Precision-Recall space for the IDTE algorithm (grouping by Tt).

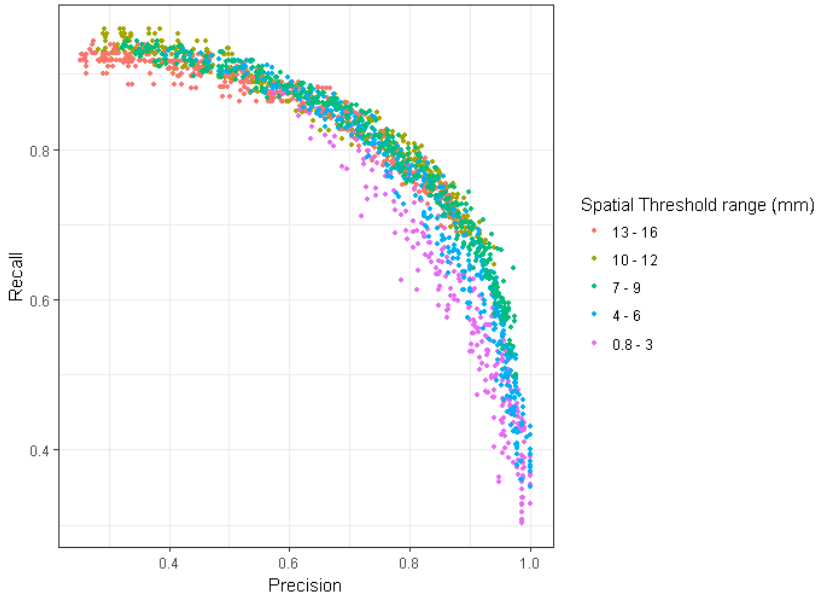


Figure C.2: Precision-Recall space for the IDTE algorithm (grouping by St).

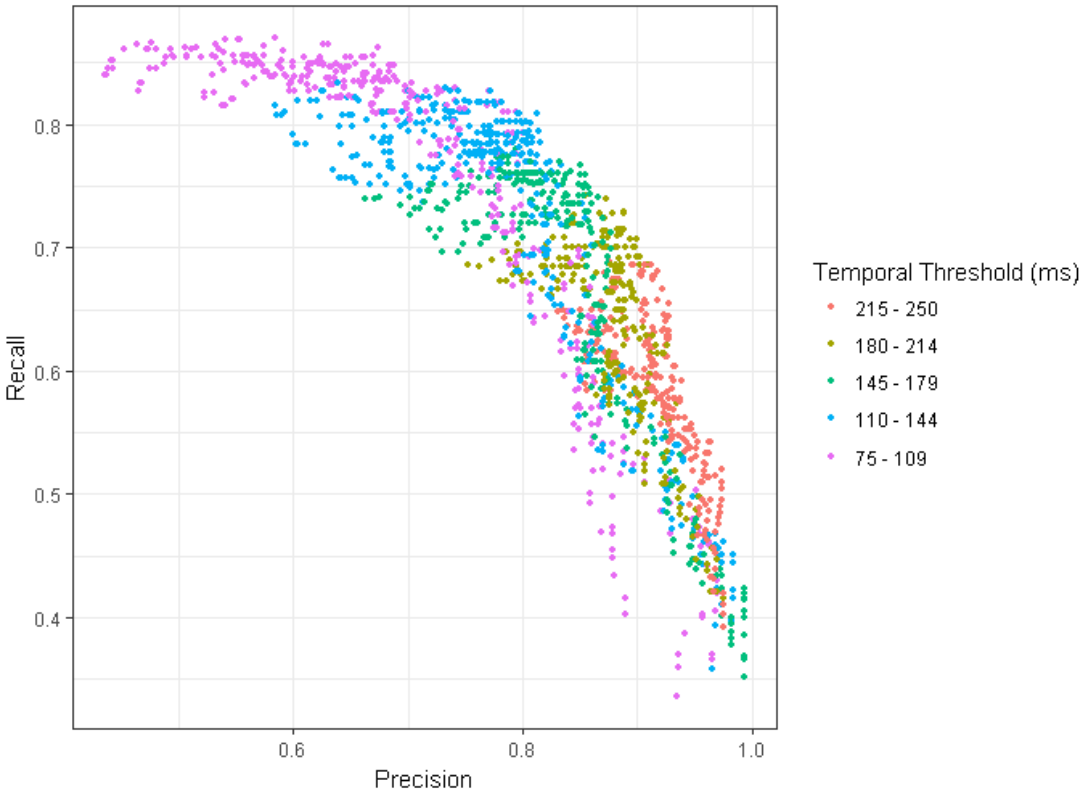


Figure C.3: Precision-Recall space for the IVT algorithm (grouping by Tt).

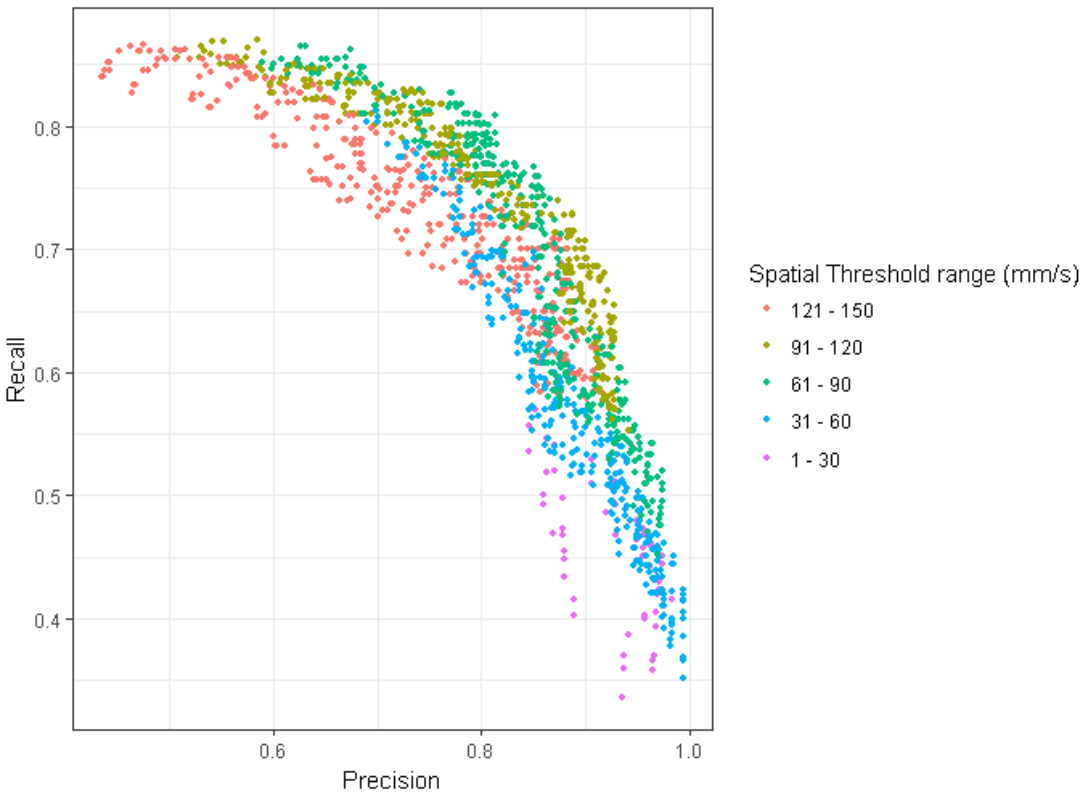


Figure C.4: Precision-Recall space for the IVT algorithm (grouping by St).

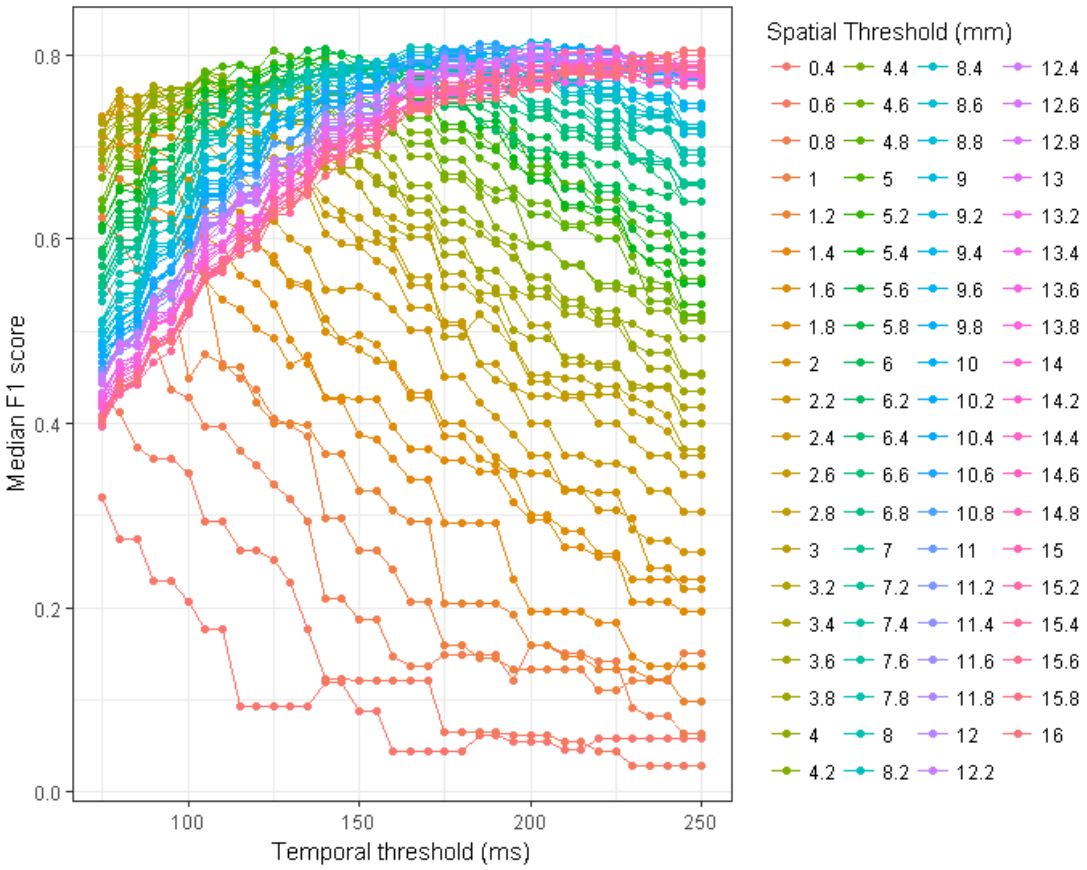


Figure C.5: F1 score for the different combinations of thresholds of the IDTE algorithm.

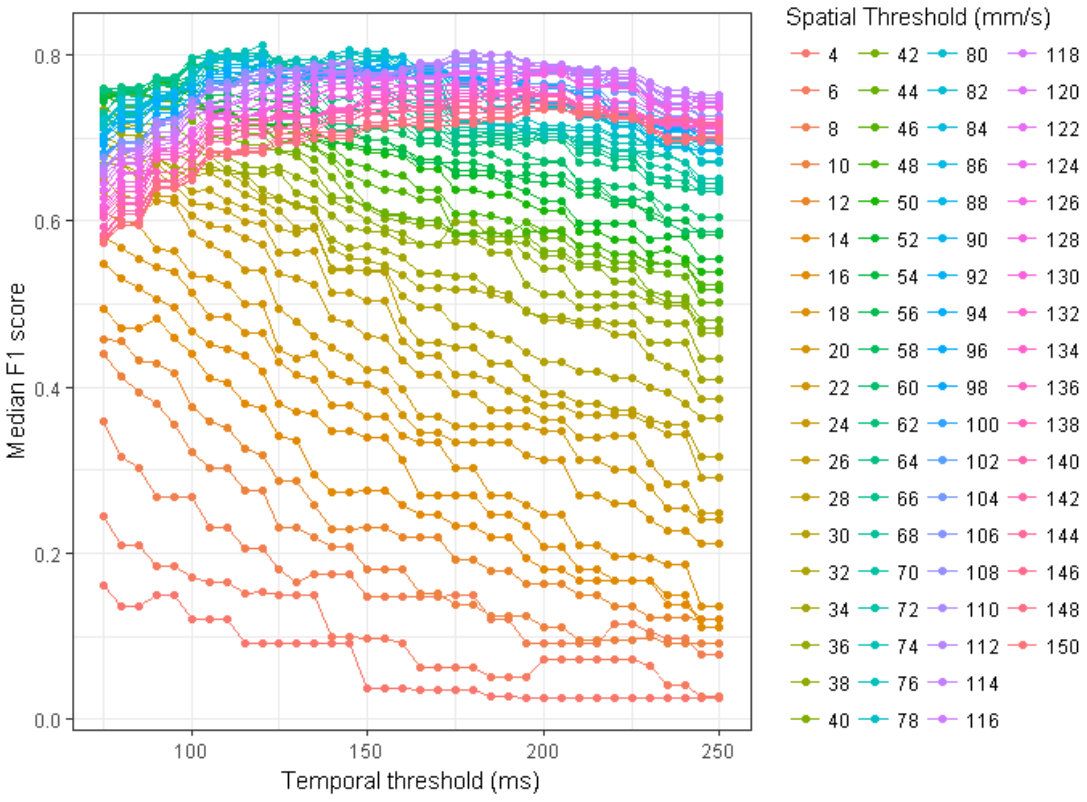


Figure C.6: F1 score for the different combinations of thresholds of the IVT algorithm.

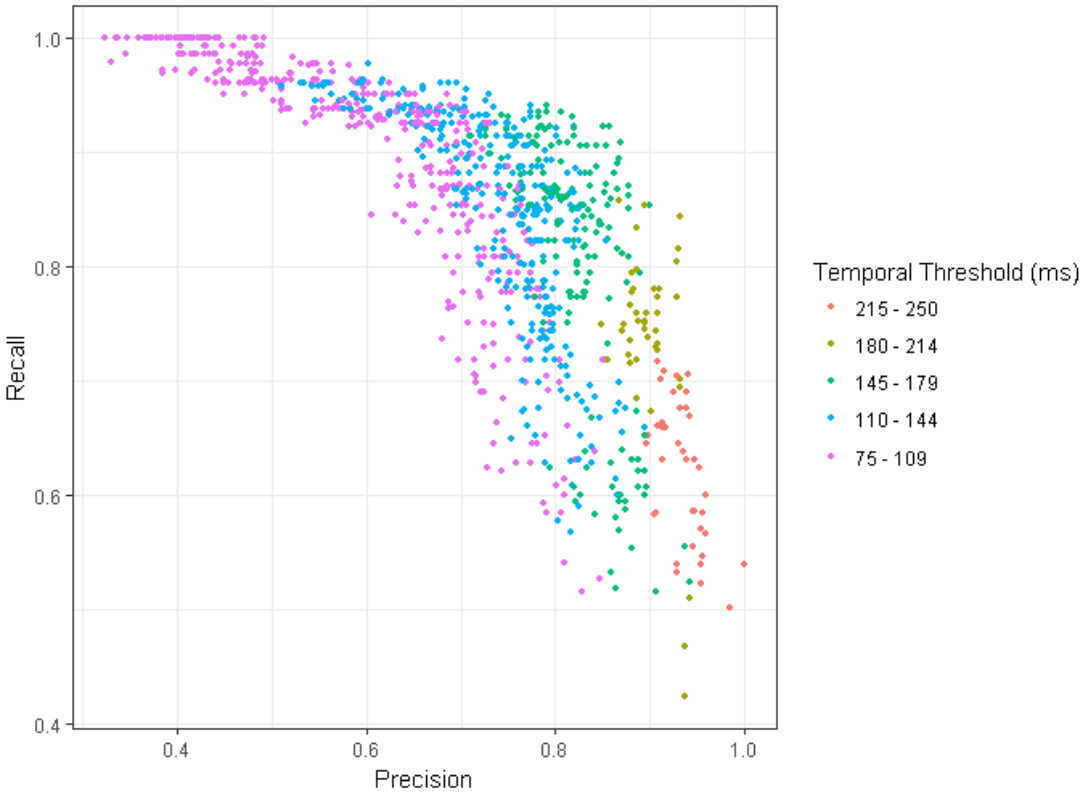


Figure C.7: IDT Precision-Recall space for the IDT algorithm (grouping by Tt, dwells only).

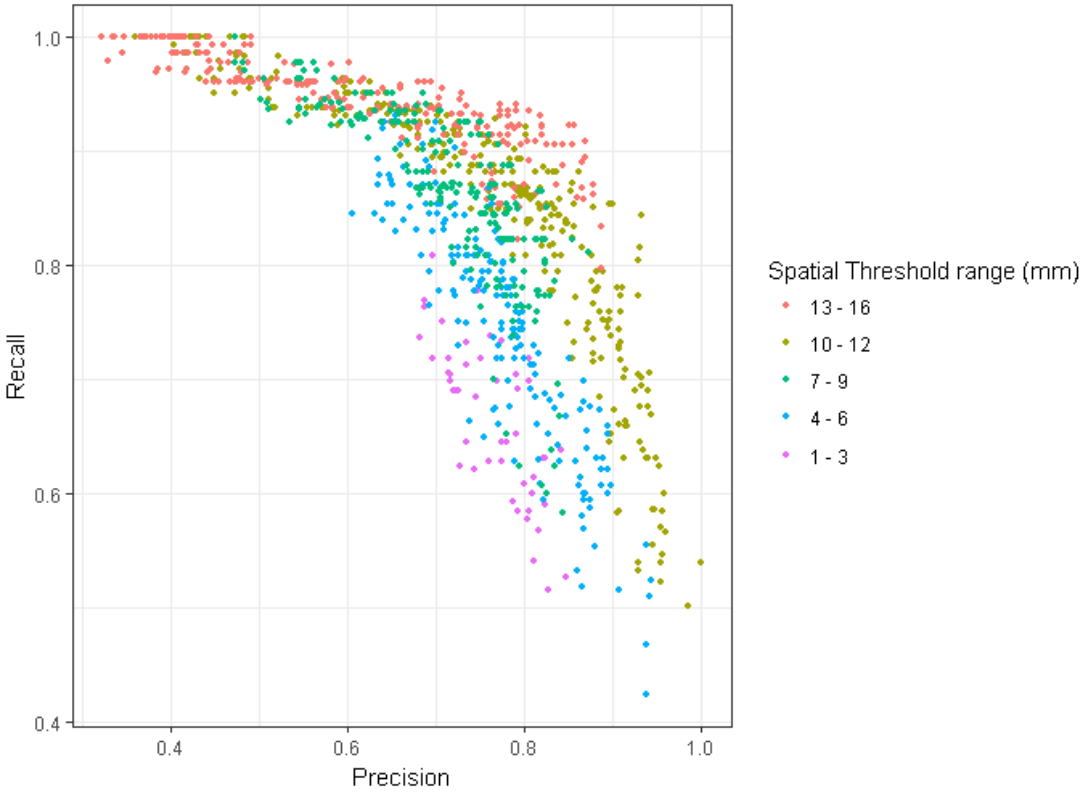


Figure C.8: Precision-Recall space for the IDT algorithm (grouping by St, dwells only).

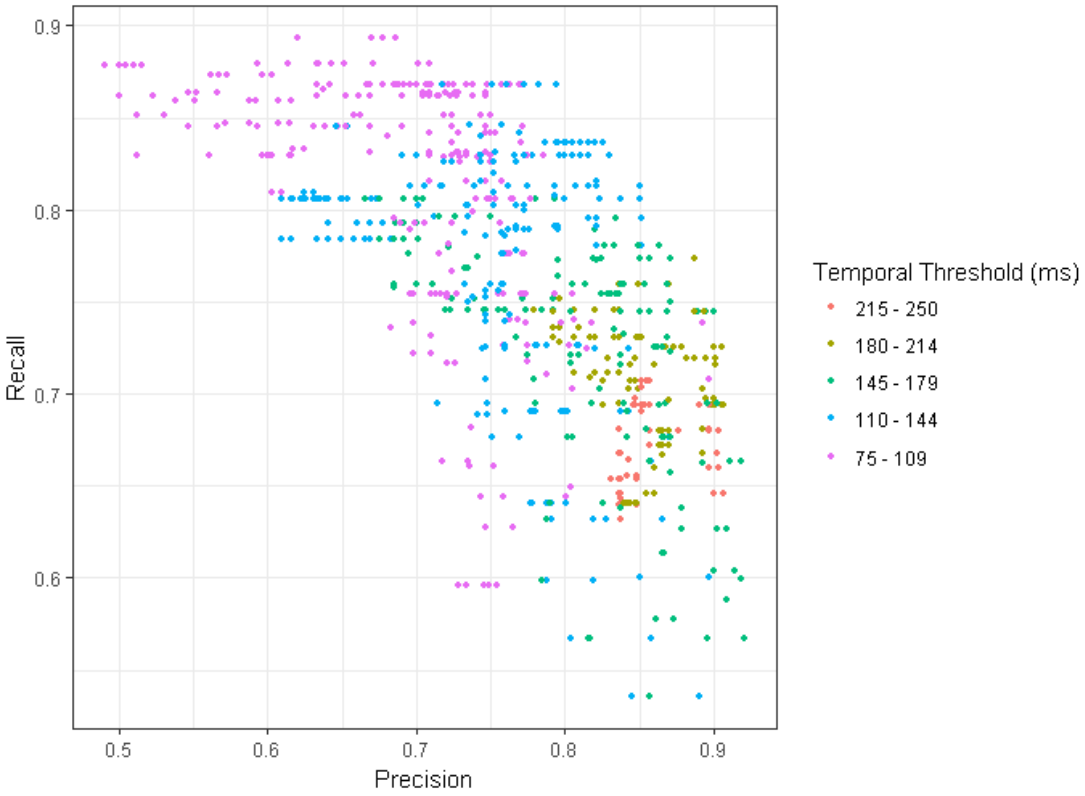


Figure C.9: Precision-Recall space for the IVT algorithm (grouping by Tt, dwells only).

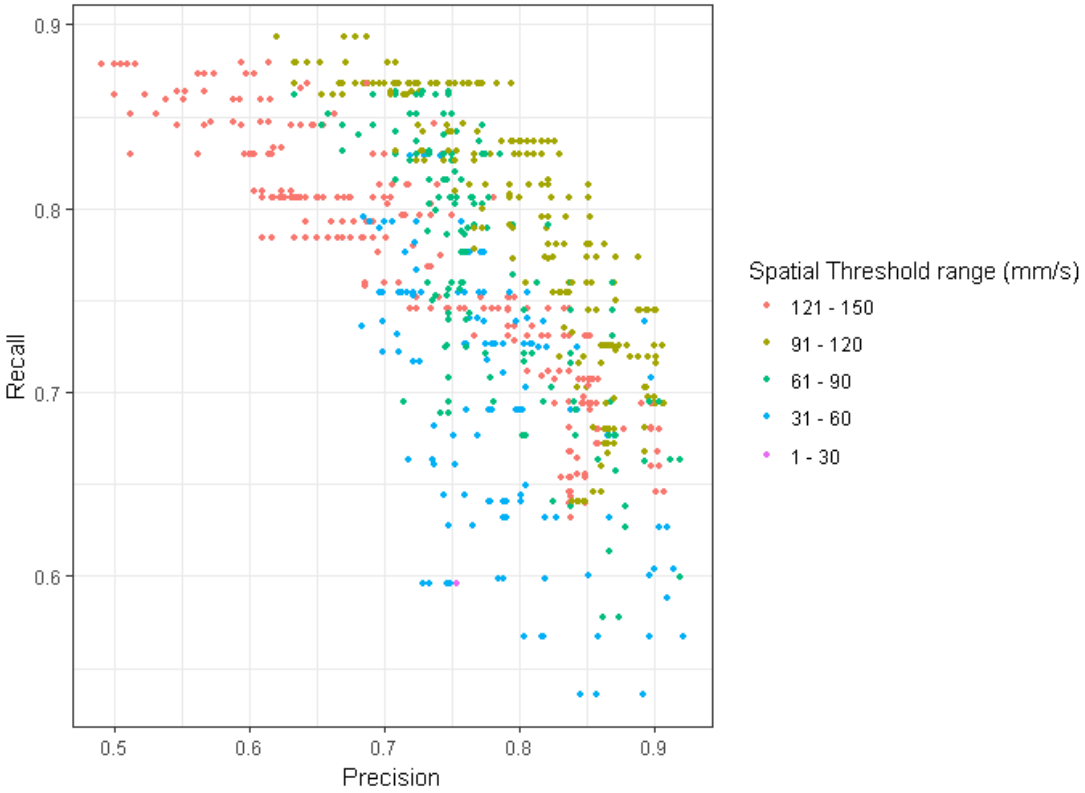


Figure C.10: Precision-Recall space for the IVT algorithm (grouping by St, dwells only).

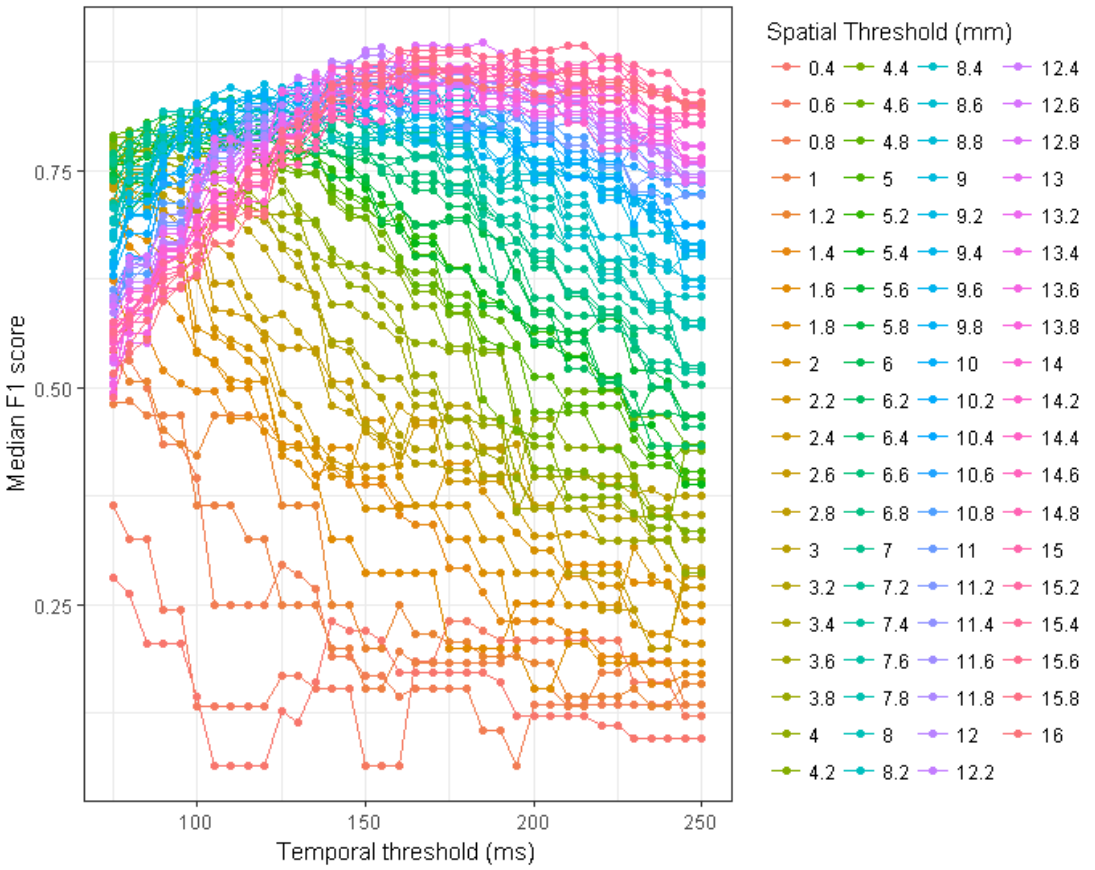


Figure C.11: F1 score for the different combinations of thresholds of the IDT algorithm (dwells only).

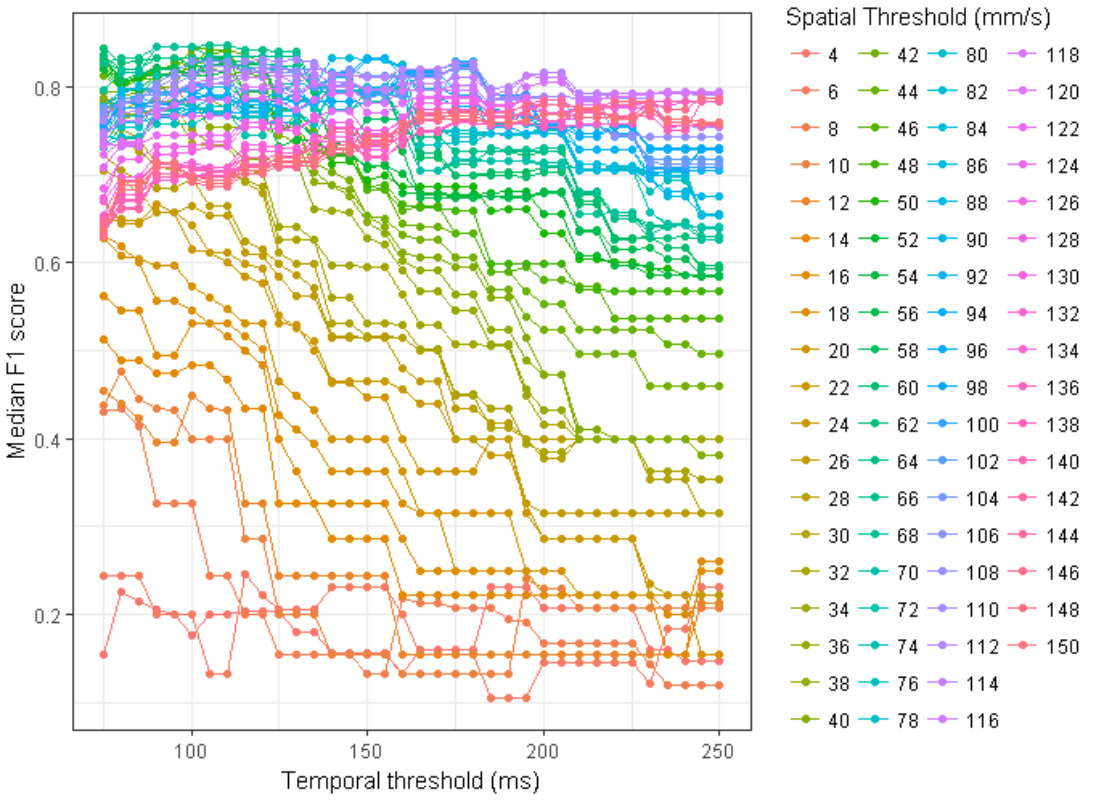


Figure C.12: F1 score for the different combinations of thresholds of the IVT algorithm (dwells only).

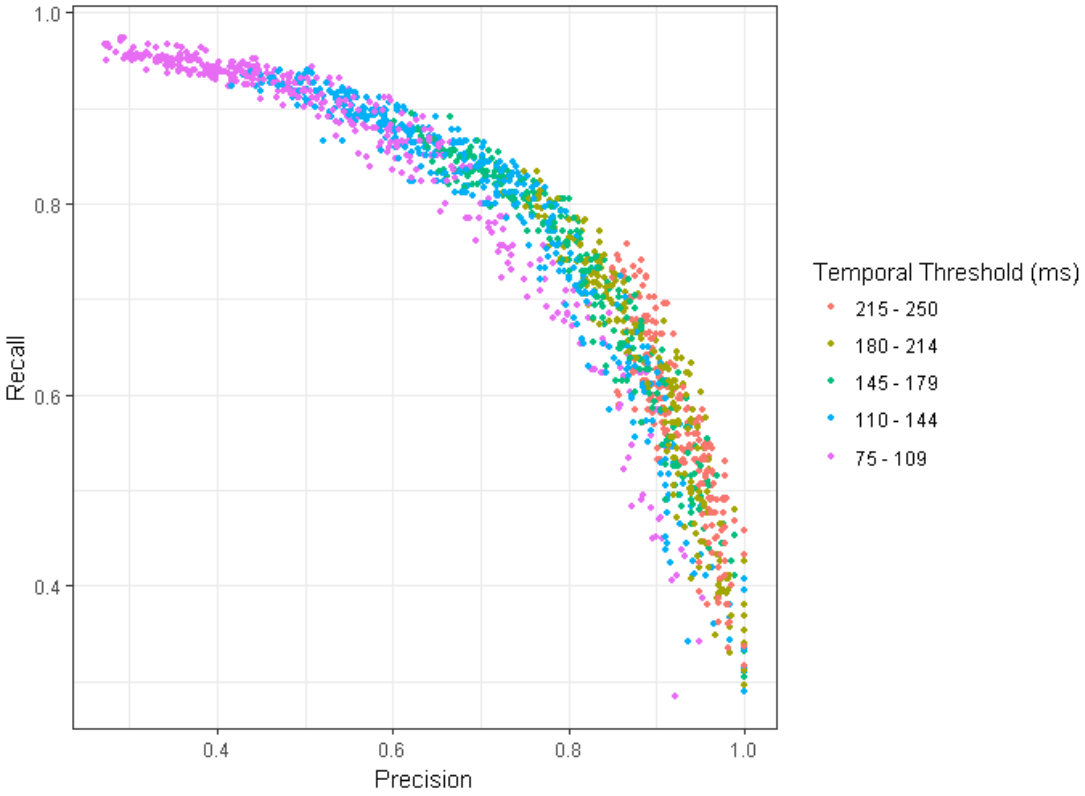


Figure C.13: IDT Precision-Recall space for the IDT algorithm (grouping by Tt, hovers only).

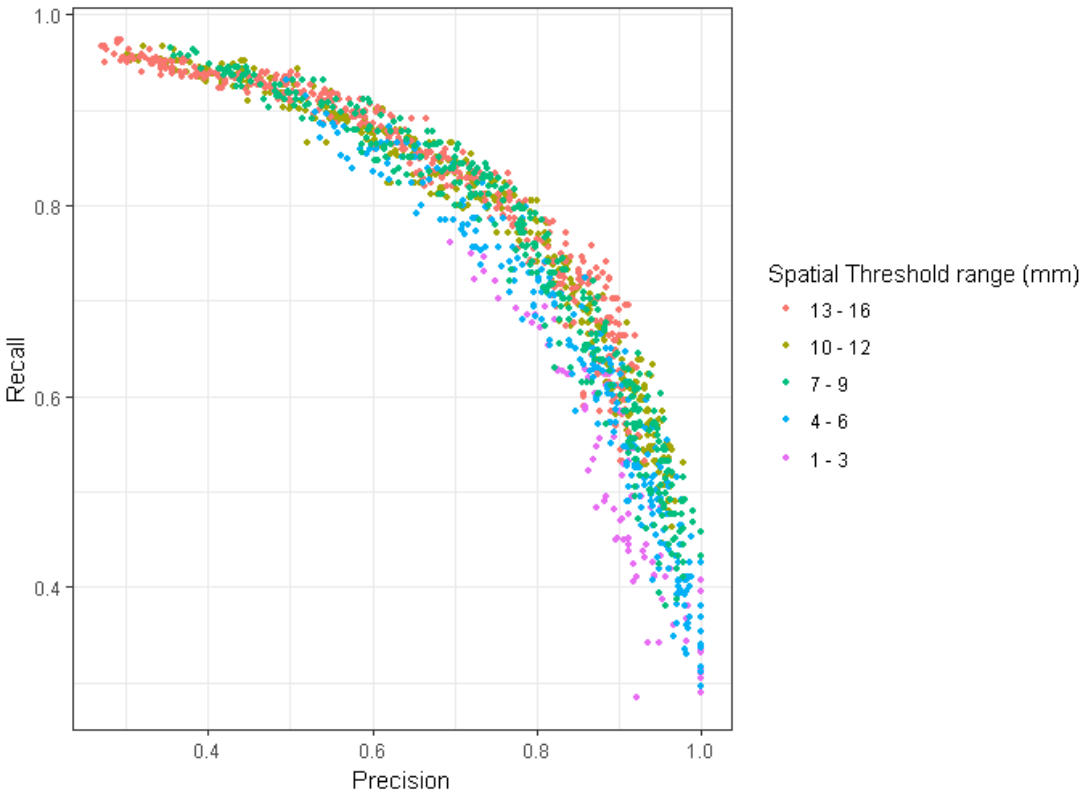


Figure C.14: Precision-Recall space for the IDT algorithm (grouping by St, hovers only).

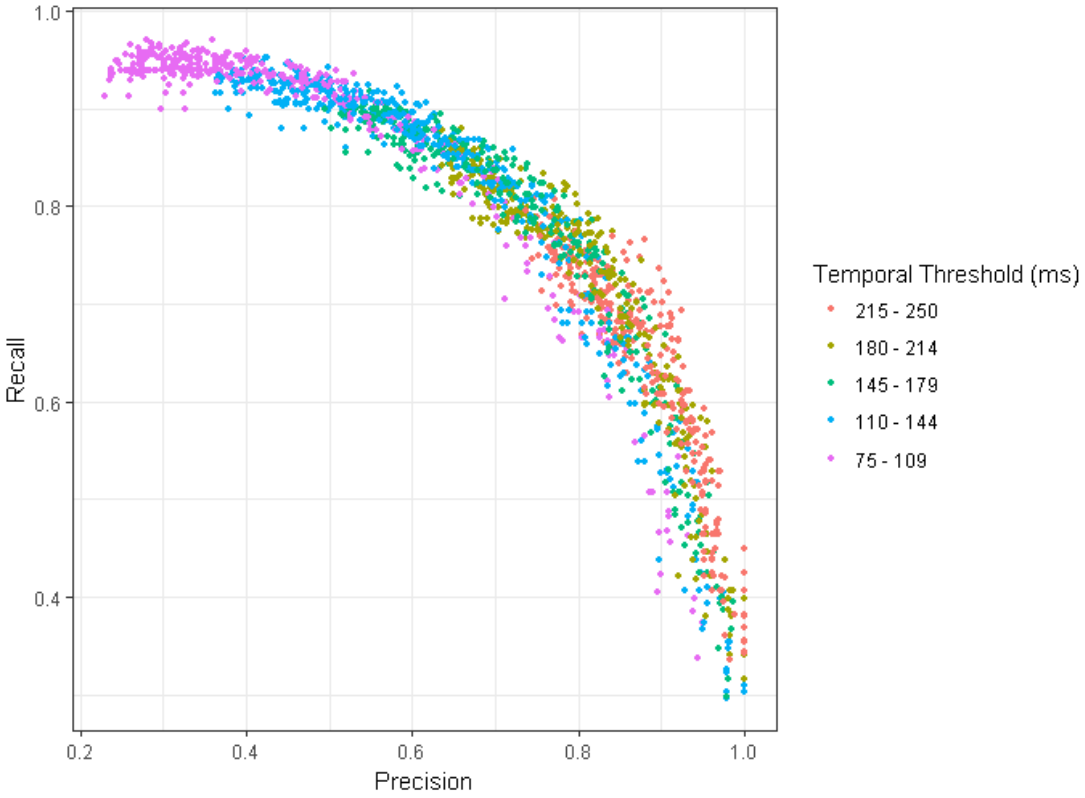


Figure C.15: Precision-Recall space for the IDTE algorithm (grouping by Tt, hovers only).

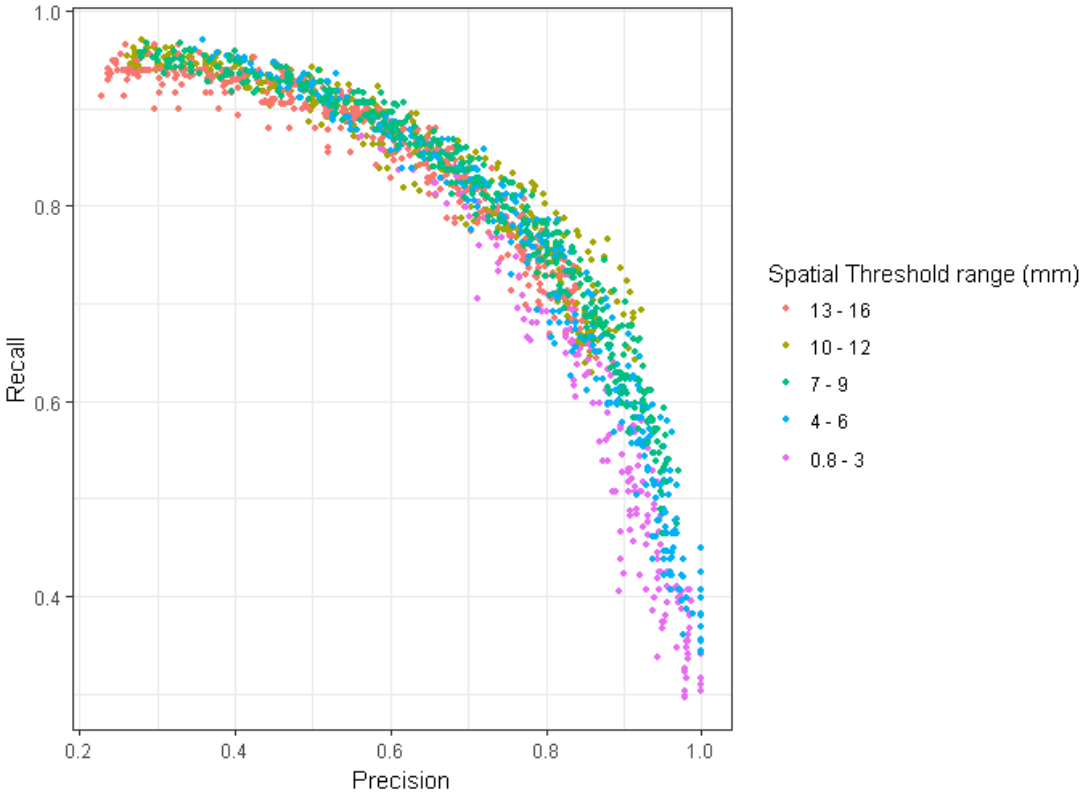


Figure C.16: Precision-Recall space for the IDTE algorithm (grouping by St, hovers only).

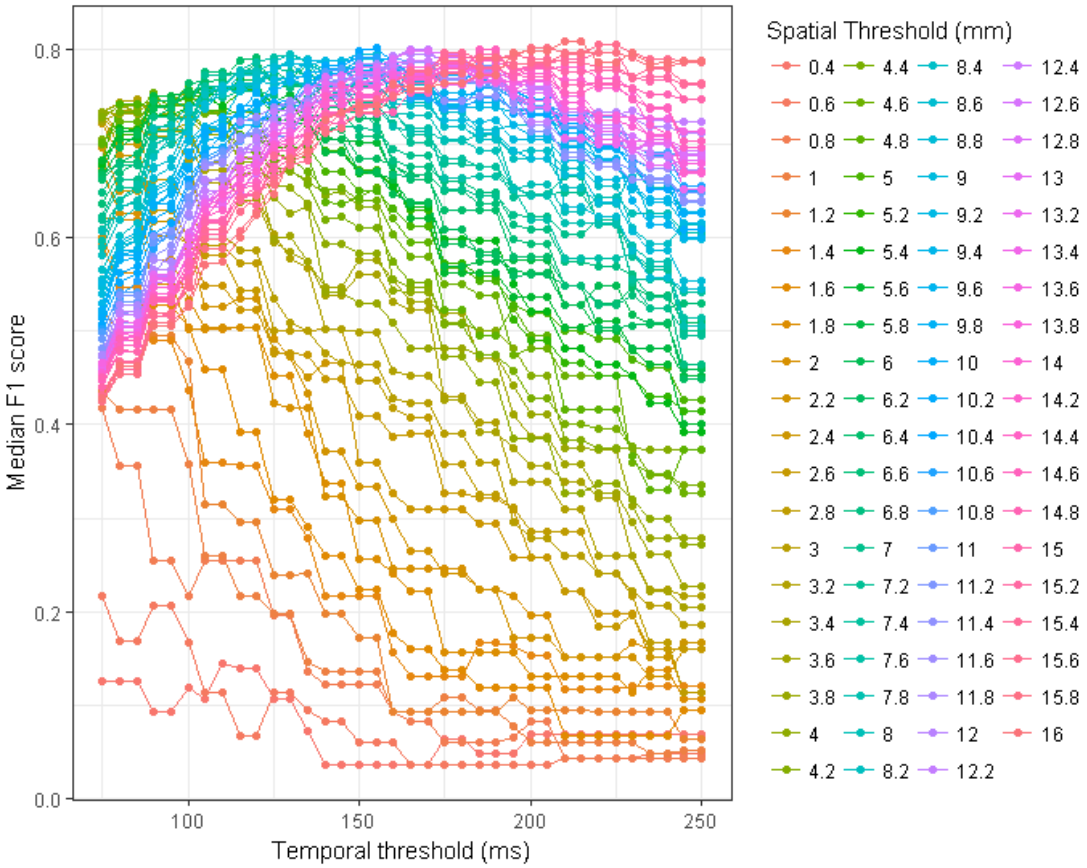


Figure C.17: F1 score for the different combinations of thresholds of the IDT algorithm (hovers only).

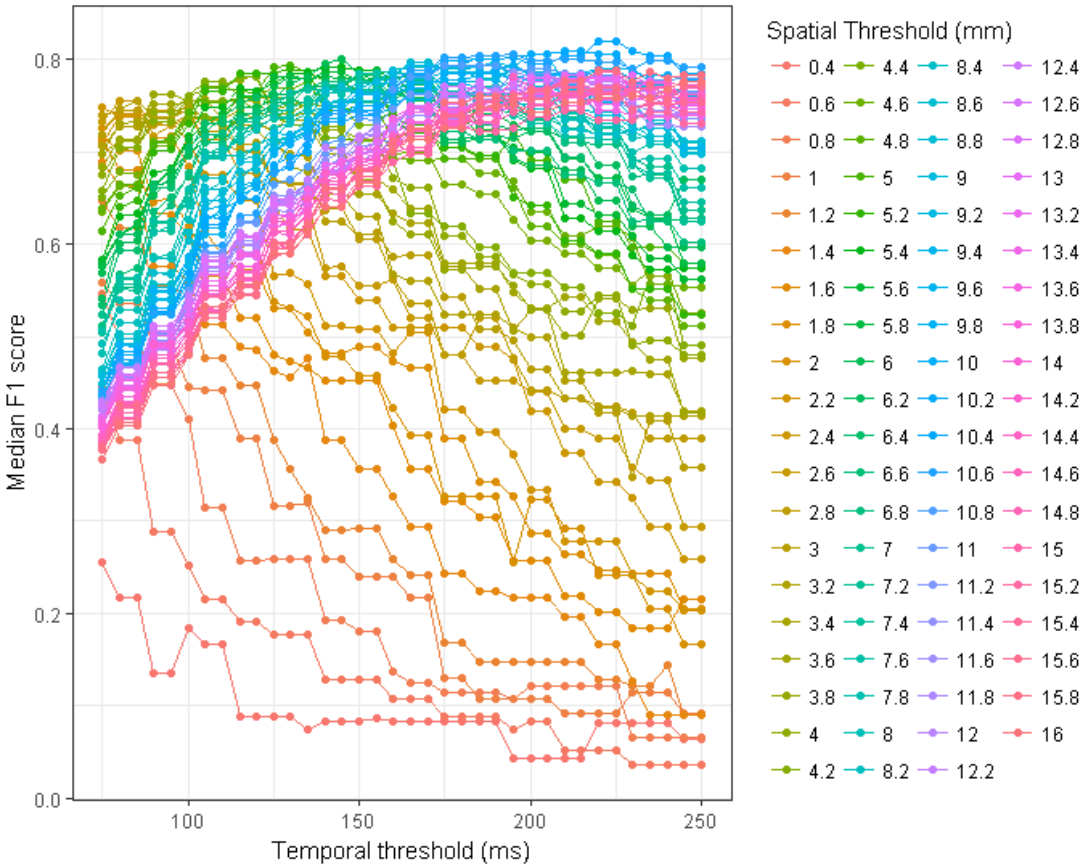


Figure C.18: F1 score for the different combinations of thresholds of the IDTE algorithm (hovers only).

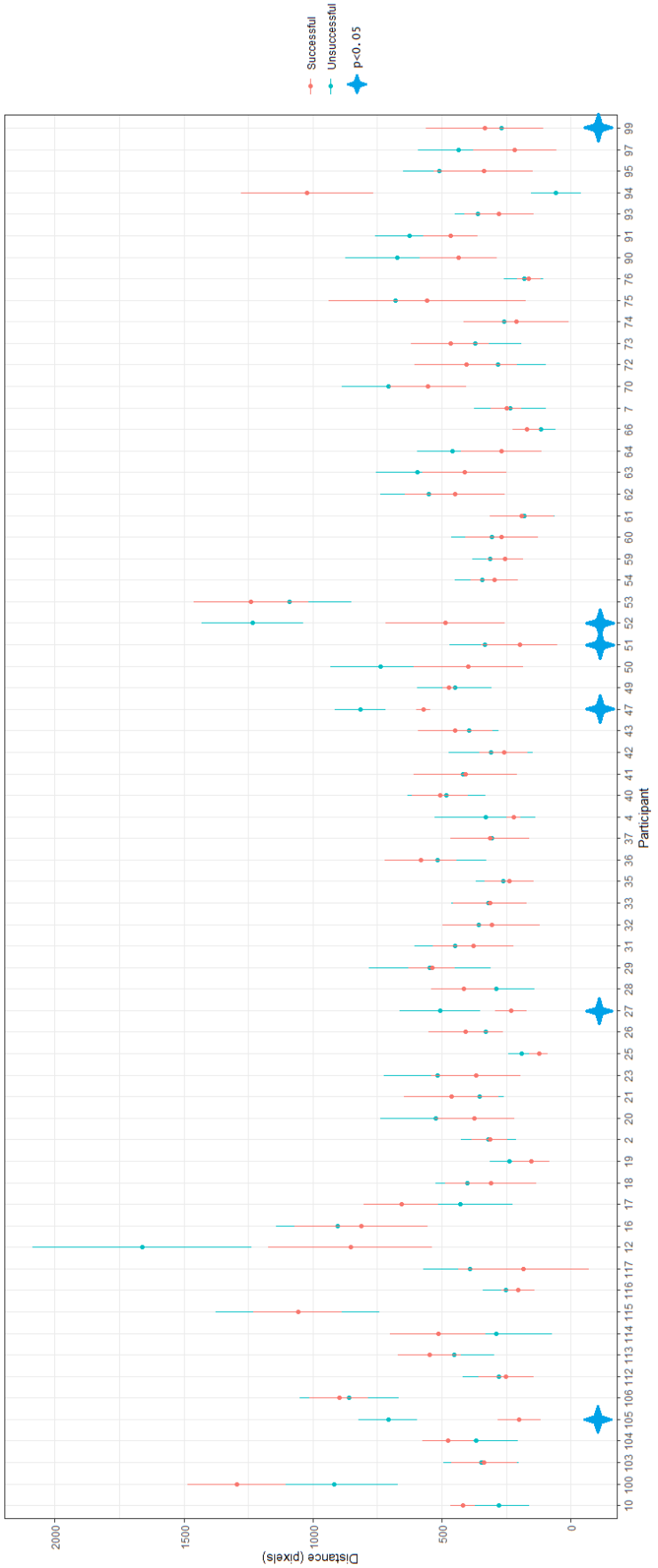


Figure C.19: Median Distance between gaze and stationary hand event (per participant).

D Gaze and Hand Motion Correlation Other Results

Table D.1: Pair-wise Z-scores for the Spearman correlation coefficients between gaze and hand comparison (X axis) per hand motion type.

Hand motion type	H-H	H-D	H-T	D-H	D-D	D-T	T-H	T-D	T-T
H-H		4.88*	3.89*	4.18*	1.65	6.17*	25.09*	26.61*	13.38*
H-D			8.44*	1.87**	4.38*	10.11*	23.54*	24.28*	15.08*
H-T				10.16*	8.96*	4.26*	38.12*	46.56*	19.83*
D-H					3.64*	13.37*	34.38*	37.02*	21.85*
D-D						14.35*	43.43*	50.76*	27.86*
D-T							43.93*	63.44*	24.76*
T-H								0.39	29.67*
T-D									44.93*

(*)p<0.01 (**)p<0.05

Table D.2: Pair-wise Z-scores for the Spearman correlation coefficients between gaze and hand comparison (Y axis) per hand motion type.

Hand motion type	H-H	H-D	H-T	D-H	D-D	D-T	T-H	T-D	T-T
H-H		1.61	18.02*	6.65*	6.64*	15.16*	10.19*	17.49*	18*
H-D			14.73*	3.48*	2.98*	12.41*	9.12*	14.04*	14.33*
H-T				29.63*	39.95*	8.37*	14.29*	3.31*	2.81*
D-H					1.26	27.48*	20.81*	30.12*	30.94*
D-D						40.44*	27.71*	44.26*	46.49*
D-T							9.48*	7.39*	9.41*
T-H								14.41*	15.9*
T-D									1.04

(*)p<0.01