Using eye tracking to study the takeover process in conditionally automated driving and piloting systems

by

Wen Ding

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering (Aeronautics)

Waterloo, Ontario, Canada, 2025

© Wen Ding 2025

Examining Committee Membership

The following served on the Examining Committee for this proposal. The decision of the Examining Committee is by majority vote.

External Member:	Huei-Yen (Winnie) Chen Associate Professor, Dept. of Industrial and Systems Engineering University at Buffalo
Supervisor(s):	Shi Cao Associate Professor, Dept. of Systems Design Engineering University of Waterloo
Supervisor(s):	Siby Samuel Assistant Professor, Dept. of Systems Design Engineering University of Waterloo
Internal-External Member:	Ewa Niechwiej-Szeedo Associate Professor, Dept. of Kinesiology and Health Science University of Waterloo
Internal Member:	Catherine Burns Professor, Dept. of Systems Design Engineering University of Waterloo
Internal Member:	Moojan Ghafurians Assistant Professor, Dept. of Systems Design Engineering University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Paper published The results of the three papers listed [38, 39] below are incorporated in Chapter 5. All three papers share the same contribution distribution, as summarized in the table below. The candidate contributed approximately 80% overall.

- Ding, W., Murzello, Y., Cao, S., Samuel, S. (2023). Where to gaze during take-over: Eye gaze strategy analysis of different situation awareness and hazard perception levels. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1), 2042-2047. https://doi.org/10.1177/21695067231193650 (Original work published 2023)
- Ding, W., Murzello, Y., Cao, S., Samuel, S. (2024). Exploring the Relationship Between Drivers' Stationary Gaze Entropy and Situation Awareness in a Level-3 Automation Driving Simulation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 68(1), 879-884. https://doi.org/10.1177/10711813241275910 (Original work published 2024)
- Ding, W., Murzello, Y., Cao, S., Samuel, S. (2025, May). Exploring the eye movement behavior between drivers with different takeover performance. Paper presented at the *CARSP/ICTCT 2025 Road Safety Conference*, Canada. Third Place, Student Paper Competition Award. Invited for journal submission to *Traffic Safety Research* (submitted).

Contributors	Research design	Data collection and analysis	Writing
Ding, W.	80%	80%	80%
Murzello, Y.		20%	
Cao, S.	10%		10%
Samuel, S.	10%		10%

Paper submitted The contribution of this part is summarized in the table below. The candidate contributed approximately 85% overall.

- The results of this paper is incorporated Chapter 5.

  Ding, W., Cao, S., Samuel, S. (2025). Visual Attention and Psychological Comfort in the Takeover Process: An Eye-Tracking Study. Submitted to *Applied Ergonomics* (under review).

Contributors	Research design	Data collection and analysis	Writing
Ding, W.	80%	100%	80%
Cao, S.	10%		10%
Samuel, S.	10%		10%

Paper in preparation Two other manuscripts are current under preparation and will be soon submitted to journals in relevant field.

- The results of this paper is incorporated Chapter 6.

  Ding, W., Cao, S., Niechwiej-szwedo, E., Samuel, S. Vigilance-Related Eye Movement Differences Between Pilots During Autopilot-Assisted Instrument Landings.

- The results of this paper is incorporated Chapter 8.

  Ding, W., Zhang, C., Cao, S., Niechwiej-szwedo, E., Samuel, S. Differences in eye movement between pilots with different flight experience during the takeover process in ILS landing.

Abstract

In a conditionally automated environment, human operators are often required to resume manual control when the autonomous system reaches its operational limits — a process referred to as takeover. This takeover process can be challenging for human operators, as they must quickly perceive and comprehend critical system information and successfully resume manual control within a limited amount of time. Following a period of autonomous control, human operators' Situation Awareness (SA) may be compromised, thus potentially impairing their takeover performance. Consequently, investigating potential approaches to enhance the safety and efficiency of the takeover process is essential. Human eyes are vital in an individual's information gathering, and eye tracking techniques have been extensively applied in the takeover studies in previous research works. The current study aims at enhancing the takeover procedure by utilizing operators' eye tracking data. The data analysis methods include machine learning techniques and the statistical approach, which will be applied to driving and piloting domains, respectively.

Simulation experiments were conducted in two domains: a level-3 semi-autonomous vehicle in the driving domain and an autopilot-assisted aircraft landing scenario in the piloting domain. In both domains, operators' eye tracking data and simulator-derived operational data were recorded during the experiments. The eye tracking data went through two categories of feature extractions: eye movement features linked predominantly to fixation and saccades, and Area-of-Interest (AOI) features associated with which AOI the gaze was located. Eye tracking features were analyzed using both traditional statistical techniques and machine learning models. Key eye tracking features included fixation-based metrics and AOI features, such as dwelling time, entry count, and gaze entropy. Operators' SA and takeover performance were measured by a series of domain-specific metrics, including Situation Awareness Global Assessment Technique (SAGAT) score, Hazard Perception Time (HPT), Takeover Time (TOT) and Resulting acceleration.

Three research topics were discussed in the current thesis and each topic included one driving study and one piloting study. In topic 1, significant differences in eye movement patterns were found between operators with higher versus lower SA, as well as between those with better and worse takeover performance. Besides the notable differences in various Area-of-Interests (AOIs) across three pre-defined Time windows (TWs), in the driving domain, drivers with a better SA and better takeover performance showed inconsistent eye movement patterns after the Takeover Request (TOR) and before they perceived hazards. In the piloting domain, pilots with shorter TOT showed more distributed and complex eye movement pattern before the malfunction alert and after resuming control. During the intervening period, their eye movements were more focused and predictable, indicating

fast identification of necessary controls with minimal visual search. In topic 2, significant differences in eye movement patterns were observed between younger and older drivers, as well as between learner and expert pilots. As for driving domain, older drivers exhibited more extensive visual scanning, indicating difficulty in effectively prioritizing information sources under time pressure. In piloting domain, expert pilots not only allocate more attention to critical instrument areas but also dynamically adjust their scanning behavior based on the current tasks. In topic 3, machine learning models trained on eye tracking features successfully performed binary classification for both SA-related and takeover performance-related metrics. Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC).

Finally, comparisons were made across Topics 1 and 2, as well as between the driving and piloting domains. The results suggest that better operators can flexibly adapt their gaze strategies to meet task demands, shifting between broad visual scanning and focused searching when appropriate. This shift in patterns underscores the importance of accounting for the specific Time window (TW) when interpreting operators' eye movements. Overall, this thesis advances the understanding of different eye movement patterns during the takeover process by exploring a range of eye tracking features. The findings support the development of operator training programs and the design of customized interfaces to enhance the safety and efficiency of takeover performance.

Acknowledgments

I would like to express my deepest gratitude to my supervisors, Prof. Shi Cao and Prof. Siby Samuel, for giving me the opportunity to begin this amazing doctoral journey at the University of Waterloo. Over the past four years, their support and guidance have meant the world to me. They not only provided incredible professional advice but also gave me the freedom to explore my own research interests. Doing a Ph.D. can be both exciting and challenging, and I truly appreciate their mentorship in both research and personal growth. Through their patience, encouragement, and dedication, they have shown me what it means to be truly great supervisors, and I feel incredibly lucky to have learned from them.

I also want to thank all the members of my thesis committee and the staff in the Department of Systems Design Engineering. This thesis and defense wouldn't have been possible without their feedback and support.

I'm very grateful to all the researchers I've had the chance to work with. Special thanks to Ms. Yovela Murzello, who graduated from our lab a few years ago. Her work on semi-autonomous vehicle provided valuable insights for my experiment and research. I also want to thank Rafael Pastorin Repato, the research coordinator at WISA. I believe that my aviation experiment wouldn't have happened without his expertise in aviation and constant support.

To my dearest friends, Kylie Ren, Qinyue Tang, and Yuan Fang – thank you for always being there for me through every high and low. Your friendship has meant so much to me. Your company and support turned Waterloo from just a place to study into a true second home, and I'll always cherish the memories we made together.

Last but not least, I want to thank my family for their endless love, encouragement, and belief in me. Even from afar, your support has been my foundation throughout this long journey; none of it would have been possible without you.

Table of Contents

Examining Committee ii

Author's Declaration iii

Statement of Contributions iv

Abstract vi

Acknowledgments viii

List of Figures xvii

List of Tables xxi

List of Abbreviations xxiv

1 Introduction 1

1.1 Research Overview 1

1.2 Research Questions and Hypotheses 2

1.2.1 Topic One 3

1.2.2 Topic Two 4

1.2.3 Topic Three 5

1.3 Research Contribution 6

1.4 Thesis Structure 8

2 Background **10**
2.1 Takeover process in different conditionally automated systems 11
2.1.1 Takeover process in driving 11
2.1.1.1 Autonomous vehicles development and taxonomy 11
2.1.1.2 Influential factors in Level 3 autonomous vehicle takeover 13
2.1.2 Takeover process in piloting 18
2.1.2.1 Autopilot takeover process 18
2.1.2.2 Instrument Landing System 20
2.2 Situation Awareness 22
2.2.1 Situation Awareness in driving and piloting research 23
2.2.2 Situation Awareness measurements 24
2.3 Eye tracking techniques 27
2.3.1 Eye tracking and Situation Awareness 30
2.3.2 Eye movement of older drivers 31

3 Experiments **33**
3.1 The driving experiment 34
3.1.1 Driving experiment design 34
3.1.2 Driving experiment instruments 37
3.1.3 Driving experiment participants 38
3.1.3.1 Participants screening criteria 38
3.1.3.2 Participants demographics and sample size 39
3.1.4 Driving experiment procedures 40
3.2 The piloting experiment 41
3.2.1 Piloting experiment design 41
3.2.2 Piloting experiment instruments 43
3.2.3 Piloting experiment participants 45
3.2.3.1 Participants screening criteria 45
3.2.3.2 Participants demographics and sample size 45
3.2.4 Piloting experiment procedures 45

4 Methodology **47**

4.1 Eye tracking data pre-processing . . . 49

4.1.1 Eye movement feature . . . 49

4.1.2 Area-of-Interest Features . . . 50

4.1.3 AOI feature calculation . . . 52

4.1.3.1 AOI feature calculation in driving . . . 54

4.1.3.2 AOI feature calculation in piloting . . . 56

4.2 Situation Awareness and Takeover performance Metrics . . . 58

4.2.1 Driving Study . . . 58

4.2.1.1 Drivers' Situation Awareness . . . 58

4.2.1.2 Drivers' takeover performance . . . 59

4.2.2 Piloting Study . . . 59

4.2.2.1 Pilots' Situation Awareness . . . 60

4.2.2.2 Pilots' takeover performance . . . 60

4.3 Time Window . . . 60

4.3.1 Driving study . . . 61

4.3.2 Piloting study . . . 62

4.4 Data processing . . . 63

4.4.1 Metrics grouping . . . 63

4.4.1.1 Driving study . . . 64

4.4.1.2 Piloting study . . . 65

4.4.2 Analytical models . . . 67

4.4.2.1 MANOVA + Welch's t-tests with FDR . . . 67

4.4.2.2 Machine Learning . . . 67

5 Differences eye movement between drivers with different Situation Awareness levels and takeover performance levels in a semi-autonomous driving system. 70

5.1 Background . . . 70
5.2 Methodology . . . 71
5.3 Results . . . 72
5.3.1 Drivers' Situation Awareness . . . 72
5.3.1.1 Dwelling time . . . 73
Results . . . 75
5.3.1.2 Entry Count . . . 77
Results . . . 79
5.3.1.3 Entropies . . . 79
Results . . . 79
5.3.2 Drivers' Takeover performance . . . 82
5.3.2.1 Dwelling time . . . 83
Results . . . 84
5.3.2.2 Entry count . . . 86
Results . . . 88
5.3.2.3 Entropy . . . 88
Results . . . 89
5.4 Discussion and conclusions . . . 91

6 Differences in eye movement between pilots with different Situation Awareness levels and takeover performance levels in an autopilot-assisted instrument landing. 93

6.1 Background . . . 93
6.2 Methodology . . . 94
6.3 Results . . . 95
6.3.1 Pilots' Situation Awareness . . . 95

6.3.1.1 Dwelling time . . . 96
Results . . . 96
6.3.1.2 Entry count . . . 98
Results . . . 98
6.3.1.3 Entropy . . . 99
Results . . . 101
6.3.2 Pilots' Takeover performance . . . 103
6.3.2.1 Dwelling time . . . 103
Results . . . 104
6.3.2.2 Entry count . . . 106
Results . . . 106
6.3.2.3 Entropy . . . 109
Results . . . 109
6.4 Discussion and conclusions . . . 112

7 Differences in eye movement between drivers from different age groups during the takeover process in a semi-autonomous driving system 114
7.1 Background . . . 114
7.2 Methodology . . . 115
7.3 Results . . . 116
7.3.1 Dwelling time . . . 117
Results . . . 117
7.3.2 Entry count . . . 120
Results . . . 120
7.3.3 Entropy . . . 123
Results . . . 125
7.4 Discussion and conclusions . . . 125

8 Differences in eye movement between pilots from different flight experience groups during the takeover process in an autopilot-assisted instrument landing. 127

8.1 Background . . . 127

8.2 Methodology . . . 128

8.3 Results . . . 129

8.3.1 Dwelling time . . . 130

Results . . . 130

8.3.2 Entry count . . . 133

Results . . . 133

8.3.3 Entropy . . . 136

Results . . . 136

8.4 Discussion and conclusions . . . 139

9 Using eye tracking features to predict drivers' Situation Awareness and takeover performance in a semi-autonomous driving system with machine learning methods. 140

9.1 Background . . . 140

9.2 Methodology . . . 141

9.3 Results . . . 142

9.3.1 Situation Awareness . . . 142

9.3.1.1 Drivers' SAGAT score prediction . . . 142

9.3.1.2 Drivers' Hazard perception prediction . . . 143

9.3.2 Takeover performance . . . 145

9.3.2.1 Drivers' Takeover time prediction . . . 145

9.3.2.2 Drivers' Resulting acceleration prediction . . . 146

9.4 Discussion and conclusions . . . 148

10 Using eye tracking features to predict pilots' Situation Awareness and takeover performance in an autopilot-assisted instrument landing with machine learning methods. 150

10.1 Background . . . 150

10.2 Methodology . . . 151

10.3 Results . . . 152

10.3.1 Situation Awareness . . . 152

10.3.1.1 Pilots' SAGAT score prediction . . . 152

10.3.2 Takeover performance . . . 153

10.3.2.1 Pilots' Takeover time prediction . . . 153

10.4 Discussion and conclusions . . . 154

11 Synthesis, Limitations, and Future Work 156

11.1 Synthesis between topics and domains . . . 156

11.1.1 Comparison between Topic 1 and 2 . . . 157

11.1.1.1 The piloting study (Comparison A.1) . . . 157

11.1.1.2 The driving study (Comparison A.2) . . . 158

11.1.2 Comparison between driving and piloting domains (Comparison B) 159

11.2 Limitations and future work . . . 160

11.3 Applications and Implications . . . 163

References 165

APPENDICES 182

A SAGAT questions 183

B Demographic Questionnaires 185

C Motion Sickness Susceptibility Questionnaire Short-form 187

D Air traffic control scripts **189**

E SAGAT questions for piloting experiment **194**

F Piloting demographic questionnaire **196**

List of Figures

2.1 J3016 driving automation taxonomy [75] . . . 12
2.2 Takeover process [56] . . . 14
2.3 Instrument Landing System [13] . . . 20
2.4 ILS Localizer . . . 21
2.5 ILS Glideslope . . . 21
2.6 ILS receiving screen . . . 21
2.7 Aircrew decision model [45] . . . 22
2.8 Number of publications using eye tracking techniques [24] . . . 27
2.9 Eye tracking glasses examples . . . 28
2.10 Fixations and Saccades [83] . . . 28
2.11 AOIs layout in a cockpit [90] . . . 29

3.1 Driving Scenarios. . . . 35
3.2 Three types of NDRTs. . . . 36
3.3 Takeover Request Icon . . . 37
3.4 Ergoneers Dikabilis 3 eye tracker . . . 38
3.5 Experiment environment setup . . . 39
3.6 Landing trajectory . . . 42
3.7 Two weather conditions in piloting experiment. . . . 43
3.8 Instrument panel with three types of failure conditions. . . . 44
3.9 Piloting experimental environment setup . . . 44

3.10 AdHawk eye tracking glasses 45
3.11 Briefing board before the formal experiment 46

4.1 Data analysis structure 47
4.2 Driving experiment AOIs 54
4.3 Driving experiment markers in the scene camera frame 55
4.4 Affine Transformation in the driving experiment 55
4.5 Piloting experiment AOIs 56
4.6 Piloting experiment markers in the scene camera frame 57
4.7 Comparison of marker visibility with and without glass layers 57
4.8 Driving experiment timelines for two types of scenarios. 61
4.9 Piloting experiment timeline. 62
4.10 SAGAT Score distribution in the driving study 64
4.11 Hazard Perception Time distribution in the driving study 64
4.12 Takeover Time distribution in the driving study 65
4.13 Resulting acceleration distribution in the driving study 65
4.14 SAGAT Score distribution in the piloting study 66
4.15 Takeover Time distribution in the piloting study 66
4.16 Machine Learning pipeline 68

5.1 Dwelling time features of drivers from two SAGAT score groups and two HPT groups across three TWs (error bars represent Standard Error of the Mean (SEM)). 76
5.2 Entry count features of drivers from two SAGAT score groups and two HPT groups across three TWs (error bars represent SEM). 78
5.3 SGE and GTE of drivers from two SAGAT score groups and two HPT groups across three TWs (error bars represent SEM). 81
5.4 Dwelling time features of drivers from two Takeover time groups and two Resulting Acceleration groups across three TWs (error bars represent SEM). 85

5.5 Entry count features of drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (error bars represent SEM). 87
5.6 SGE and GTE of drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (error bars represent SEM). 90

6.1 Dwelling time features of pilots from two SAGAT score groups across three TWs (error bars represent SEM). 97
6.2 Entry count features of pilots from two SAGAT score groups across three TWs (error bars represent SEM). 100
6.3 SGE and GTE of pilots from two SAGAT score groups across three TWs (error bars represent SEM). 102
6.4 Dwelling time features of pilots from two Takeover Time groups groups across three TWs (error bars represent SEM). 105
6.5 Entry count features of pilots from two Takeover Time groups groups across three TWs (error bars represent SEM). 108
6.6 SGE and GTE of pilots from two Takeover time groups across three TWs (error bars represent SEM). 111

7.1 Dwelling time features of drivers from two age groups across three TWs (error bars represent SEM). 119
7.2 Entry count features of drivers from two age groups across three TWs (error bars represent SEM). 122
7.3 SGE and GTE of drivers from two age groups across three TWs (error bars represent SEM). 124

8.1 Dwelling time features of pilots from two experience groups across three TWs (error bars represent SEM). 132
8.2 Entry count features of pilots from two experience groups across three TWs (error bars represent SEM). 135
8.3 SGE and GTE of pilots from two experience groups across three TWs (error bars represent SEM). 138

9.1 Confusion matrix of the RBF SVM model in classifying higher vs. lower SAGAT scores 143

9.2 Confusion matrix of the Gaussian Process model in classifying shorter vs. longer hazard perception time 144
9.3 SHAP values ranking of the top 9 eye tracking features used in the HPT classification task. 145
9.4 Confusion matrix of the RBF SVM model in classifying shorter vs. longer Takeover time 147

10.1 Confusion matrix of the Random Forest model in classifying higher vs. lower SAGAT scores 153
10.2 Confusion matrix of the Random Forest model in classifying shorter vs. longer takeover time 155

11.1 Comparisons between results from two topics and two domains 157

D.1 ATC Dialogue Timing in Relation to the Landing Trajectory 189

List of Tables

2.1 SA measurements 26
2.2 Eye tracking features 30

3.1 Participants' age groups 40
3.2 Participants' experience group 46

4.1 Eye movement features 50
4.2 AOI feature 51
4.3 Gaze-to-AOI Table using One-Hot Encoding 52
4.4 Gaze Transition Table 52
4.5 Evaluated Metrics for SA and Takeover Performance in the driving experiment 58
4.6 Evaluated Metrics for SA and Takeover Performance in the piloting experiment 60

5.1 T-tests for Dwelling times features between drivers from two SAGAT score groups and two HPT groups across three TWs (* indicates $p < .05$) 74
5.2 T-tests for Entry count features between drivers from two SAGAT score groups and two HPT groups across three TWs (* indicates $p < .05$) 77
5.3 T-tests for SGE and GTE between drivers from two SAGAT score groups and two HPT groups across three TWs (* indicates $p < .05$) 80
5.4 T-tests for Dwelling Time features between drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (* indicates $p < .05$) 83

5.5 T-tests for Entry count features between drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (* indicates $p < .05$) . . . 86
5.6 T-tests for SGE and GTE between drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (* indicates $p < .05$) 89

6.1 T-tests for Dwelling time features between pilots from two SAGAT score groups across three TWs (* indicates $p < .05$) . . . 97
6.2 T-tests for Entry count features between pilots from two SAGAT score groups across three TWs (* indicates $p < .05$) . . . 98
6.3 T-tests for SGE and GTE features between pilots from two SAGAT score groups across three TWs (* indicates $p < .05$) . . . 99
6.4 T-tests for Dwelling time features between pilots from two Takeover time groups across three TWs (* indicates $p < .05$) . . . 104
6.5 T-tests for Entry Count features between pilots from two Takeover time groups across three TWs (* indicates $p < .05$) . . . 106
6.6 T-tests for SGE and GTE features between pilots from two Takeover Time groups across three TWs (* indicates $p < .05$) . . . 109

7.1 T-tests for Dwelling times features between drivers from two age groups across three TWs (* indicates $p < .05$) . . . 118
7.2 T-tests for Entry count features between drivers from two age groups across three TWs (* indicates $p < .05$) . . . 120
7.3 T-tests for SGE and GTE between drivers from two age groups across three TWs (* indicates $p < .05$) . . . 123

8.1 T-tests for Dwelling time features between pilots from two experience groups across three TWs (* indicates $p < .05$) . . . 131
8.2 T-tests for Entry count features between pilots from two experience groups across three TWs (* indicates $p < .05$) . . . 133
8.3 T-tests for SGE and GTE between pilots from two experience groups across three TWs (* indicates $p < .05$) . . . 136

9.1 Classification performance of five models on the drivers' SAGAT score . . . 142

9.2 Classification performance of five models on the drivers' Hazard Perception time 144
9.3 Classification performance of five models on the drivers' Takeover time . . 146
9.4 Classification performance of five models on the resulting acceleration . . . 147

10.1 Classification performance of five models on the pilots' SAGAT scores . . . 152
10.2 Classification performance of five models on the pilots' takeover time . . . 154

List of Abbreviations

ADS Automated Driving System 13

AFCS Automated Flight Control System 18, 93, 127, 162

AGL above ground level 42, 43

AOI Area-of-Interest vi, 4, 5, 7, 8, 44, 48–55, 57, 67, 71–75, 77, 82, 83, 86, 89, 91, 94–96, 98, 103, 106, 112, 113, 115–117, 120, 128–130, 133, 141, 145, 151, 161, 163

AOIs Area-of-Interests vi, 2, 7, 28–30, 50–56, 71–74, 77, 79, 80, 83, 84, 86, 88, 91, 96, 98, 99, 103, 104, 106, 107, 109, 110, 112, 115–117, 120, 125, 130, 133, 136, 137, 141, 145, 151, 157–159, 161, 163

ASL above sea level 41

ATC Air Traffic Controller 21, 42, 46, 189, 194

AUC Area Under the ROC Curve vii, 68, 141, 142, 144, 146–148, 151, 152, 154

CDI Cockpit Display Interfaces 26

CTT Critical Tracking Task 16, 35, 36

CYKF Region of Waterloo International Airport 41, 42

DDT Dynamic Driving Task 13

DH Decision Height 21

FAF final approach fix 42

FSD Full Self-Driving 11

GPC Gaussian Process Classifier 68, 145

GS Glideslope 20, 21

GTE Gaze transition entropy 29, 50, 51, 53, 67, 71, 79, 80, 88, 89, 91, 94, 99, 101, 109, 110, 112, 115, 123, 125, 126, 128, 136, 137, 139, 141, 151, 157–159, 163

HPT Hazard Perception Time vi, 17, 32, 59, 64, 72–75, 77, 79, 80, 91, 141–145, 148

HR Heart rate 25, 26

HRV Heart rate variability 25, 26

ICAO International Civil Aviation Organization 20, 21

IFR Instrumental Flight Rules 25

ILS Instrument Landing System 6–8, 20, 21, 45, 46, 128, 151, 162, 163

LightGMB Light Gradient Boosting Machine 30

LOC Localizer 20, 21, 41–43, 95, 129

MANOVA Multivariate Analysis of Variance 3, 4, 48, 67, 71, 72, 82, 94, 95, 103, 116, 129

NDRT Non-driving Related Task 34, 35, 72, 116

NDRTs Non-driving Related Tasks 13–17, 23, 32, 38, 40, 71, 140

ODD Operational Design Domain 13, 14, 70

PPL Private Pilot License 45

RF Random Forest 68

RVRs Runway visual ranges 21

SA Situation Awareness vi, vii, 1–10, 20, 22–27, 30–34, 37, 47–49, 58–65, 67, 68, 71–73, 77, 79, 91, 92, 94–96, 98, 99, 112, 126, 128, 140, 141, 148, 150, 151, 154, 155, 161–163

SACRI Situation Awareness Control Room Inventory 24, 26

SAE Society of Automotive Engineers 12, 13, 70

SAGAT Situation Awareness Global Assessment Technique vi, 17, 24–26, 58–60, 63–66, 72–75, 77, 79, 80, 91, 95, 96, 98, 99, 101, 112, 141–143, 148, 149, 151–154, 161, 162, 183, 194

SART Situation Awareness Rating Technique 25, 26

SEM Standard Error of the Mean xviii, xix, 74, 76–79, 81, 84–87, 90, 96–100, 102, 104–106, 108, 109, 111, 117, 119, 120, 122–124, 130, 132, 133, 135, 136, 138

SGE Stationary gaze entropy 29, 30, 50, 51, 53, 67, 71, 79, 88, 89, 94, 99, 101, 109, 110, 112, 115, 123, 125, 126, 128, 136, 137, 139, 141, 151, 157, 158

SPAM Situation Present Assessment Method 25, 26

SuRT Surrogate Reference Task 16, 35, 36

SVM Support Vector Machine 68

TO/GA takeoff/go-around 19

TOR Takeover Request vi, 6, 13–17, 23, 36, 37, 40, 54, 59, 60, 72, 79, 80, 84, 88, 91, 115–117, 120, 121, 125, 158

TORs Takeover Requests 5, 16

TOT Takeover Time vi, 14, 15, 18, 32, 43, 59, 60, 64–66, 82–84, 86, 88, 89, 91, 92, 103, 104, 106, 107, 109, 112, 133, 136, 141, 142, 145, 146, 148, 149, 151–155, 158, 159

TW Time window vii, 7–9, 47, 48, 61, 62, 67, 71–92, 95–100, 102–112, 116, 117, 119–122, 124, 125, 129–139, 141–143, 145, 146, 151–153, 156–160, 163

TWs Time windows vi, 3, 4, 8, 37, 43, 60–63, 67, 71–75, 77, 79, 82–84, 86, 88, 91, 94–96, 98, 99, 101, 103, 104, 106, 109, 112, 113, 115–117, 120, 123, 125, 128–130, 133, 136, 139, 141–143, 146, 151–153, 156–159, 163, 164

UAV Unmanned Aerial Vehicle 18

VFR Visual Flight Rules 25

WISA Waterloo Institute for Sustainable Aeronautics 43

Chapter 1

Introduction

1.1 Research Overview

As automated control systems continue to advance, the need for continuous manual operation by human operators has diminished. Automation increasingly handles routine tasks, allowing operators to redirect their cognitive resources to higher-level monitoring or decision-making activities. However, despite these advancements, most autonomous systems still require supervision and intervention from humans in certain situations. For example, when facing ambiguous or degraded road lane markings, camera-based perception systems in autonomous vehicles may struggle to reliably plan the driving trajectory. Even if modern systems increasingly integrate complementary sensors such as LiDAR and radar to overcome such challenges, lane ambiguity still poses difficulties for vision-dominant approaches. Likewise, if the radar sensor of the autopilot system experiences an anomaly, the pilot is compelled to take over the control manually. This shift from automated to manual control, commonly referred to as the takeover process, can pose significant challenges to operators. Operators may struggle to collect requisite contextual information and regain SA due to limited exposure to the evolving context during automation. As a result, they may have difficulty determining the appropriate actions and need to exert greater cognitive effort to meet the original task objectives, potentially increasing the risk of errors or accidents. Furthermore, operators with different experiences have different strategies for gathering information. Proficient individuals gather information more efficiently, while those with comparatively lower competence might need a longer time to access the same information.

To improve the efficiency and safety of the takeover process, researchers have conducted

extensive research on studying operators SA and takeover performance. Operators' SA is a critical cognitive consciousness in the takeover process [30]. SA includes the extent to which operators perceive information from their surroundings, subsequently comprehending and projecting it into subsequent actions [44]. Operators' takeover performance typically refers to their reaction time in response to visual or auditory takeover alerts, as well as the quality of their manual control following the takeover.

In the context of human operation within complex systems, such as driving or piloting, a large portion of contextual information is acquired through human eyes [61]. Eye tracking has proven to be a substantiated and pertinent technique for monitoring human attention allocation and cognitive processes. In previous research, various methodologies have been formulated to find the relevance between eye tracking gaze data and the SA [63, 94, 153] as well as takeover performance [155]. However, previous work mainly focuses on using statistical analysis to identify characteristics of eye movement at the level of individual trials, ignoring the temporal variations of eye tracking features within a trial. Moreover, limited attention has been given to the use of machine learning models that can predict quantified levels of SA and takeover performance by leveraging detailed eye-tracking features derived from temporal eye movement patterns and gaze distributions across different AOIs.

The present thesis aimed to improve the safety and efficiency of the takeover process in two fields: driving and piloting, with eye tracking techniques. The central focus is exploring the use of various categories of eye tracking features. On the one hand, statistical analyses were conducted to examine how operators from different SA, performance, and demographic groups exhibited distinct eye movement patterns. On the other hand, machine learning techniques were applied to the eye tracking features to predict qualitative measures of operators' performance, including their SA and takeover performance. To accomplish this, two experiments were conducted, including one driving experiment and one piloting experiment, each containing the integration of eye tracking devices. Following the experiments, the collected data was used to explore topics one, two, and three below. In each topic, the same research idea was applied to both the driving and piloting domains, and the same data analysis pipeline was used on the driving and piloting data, respectively.

1.2 Research Questions and Hypotheses

Based on previous discussion, this thesis is organized around three research topics, each of them including a driving study and a piloting study that have shared research idea and methodology. Within each Topic, the research question is discussed, followed by the corresponding hypotheses.

1.2.1 Topic One

The Topic 1 is: **Differences in eye movement between operators with different SA levels and takeover performance levels in a conditionally automated system.**

This topic aims to investigate weather there are significant differences in the operators' eye movement between the higher versus the lower SA group, as well as between those with higher versus lower takeover performance group. The analysis is conducted in both domains:

- **Drive study:** Differences in eye movement between drivers with different SA levels and takeover performance levels in a semi-autonomous driving system.
- **Piloting study:** Differences in eye movement between pilots with different SA levels and takeover performance levels in an autopilot-assisted instrument landing.

Research Question: How does operators with different SA levels and takeover performance levels move their eyes differently in a conditionally automated system?

Hypothesis: It is hypothesized that operators' eye-movement behavior would differ depending on their SA and takeover performance levels, within each of the three TWs of the takeover process. To test his, a Multivariate Analysis of Variance (MANOVA) was conducted to examine differences in eye movement patterns between the higher versus the lower SA group, as well as between those with higher versus lower performance group. Follow-up univariate Welch's t-tests were then applied to individual eye-tracking features to identify which specific feature contributed to the overall effects. Based on these analyses, the more detailed hypotheses are formulated as follows:

Multivariate Hypothesis:

- $H_0(multivariate)$: There is no difference in the combined multivariate profile of eye-movement features between groups (either between SA groups or between takeover performance groups).
- $H_a(multivariate)$: At least one linear combination of the eye-movement features differs significantly between groups (either between SA groups or between takeover performance groups).

Follow-up Univariate Hypotheses (if multivariate test is significant): For each eye-movement feature, univariate tests were conducted using Welch's t-tests with Benjamini–Hochberg false discovery rate (FDR) correction to control for multiple comparisons.

- $H_0(univariate)$: The mean feature value does not differ between groups (either between SA groups or between takeover performance groups).
- $H_a(univariate)$: The mean feature value differs between groups (either between SA groups or between takeover performance groups).

The outcome of this study can intuitively show how operators from different SA and performance groups move their eyes differently, and this is crucial for making explainable conclusions to ensure safety. Furthermore, the conclusions can be used to instruct on operators' information-gathering strategies in the future.

1.2.2 Topic Two

The Topic 2 is: Differences in eye movement between operators from different population groups during the takeover process in a conditionally automated system.

The purpose of this topic is to find out if there are significant differences in operators' eye movement between different population groups. This topic explored the same AOI-related features as topic 1. The analysis is conducted in both domains:

- **Drive study:** Differences in eye movement between drivers from different age groups during the takeover process in a semi-autonomous driving system.
- **Piloting study:** Differences in eye movement between pilots from different flight experience group during the takeover process in an autopilot-assisted instrument landing.

Research Question: How does operators from different population groups move their eyes differently during the takeover process in a conditionally automated system?

Hypothesis: It is hypothesized that operators' eye movement patterns would differ depending on their background within each of the three TWs of the takeover process. To test his, a MANOVA was conducted to examine differences in eye movement patterns between drivers from different age groups and pilots from different experience groups. Follow-up univariate Welch's t-tests were then applied to individual eye-tracking features to identify which specific feature contributed to the overall effects. Based on these analyses, the more detailed hypotheses are formulated as follows:

Multivariate Hypothesis:

- $H_0(multivariate)$: There is no difference in the combined multivariate profile of eye-movement features between groups (either between age groups or experience groups).
- $H_a(multivariate)$: At least one linear combination of the eye-movement features differs significantly between groups (either between age groups or experience groups).

Follow-up Univariate Hypotheses (if multivariate test is significant): For each eye-movement feature, univariate tests were conducted using Welch's t-tests with Benjamini–Hochberg false discovery rate (FDR) correction to control for multiple comparisons.

- $H_0(univariate)$: The mean feature value does not differ between groups (either between age groups or experience groups).
- $H_a(univariate)$: The mean feature value differs between groups (either between age groups or experience groups).

The outcome of this study demonstrated how operators from different population groups move their eyes differently, which can be used to reformulate the Takeover Requests (TORs) or system alert based on group-specific characteristics.

1.2.3 Topic Three

The Topic 3 is: Using eye tracking features to predict operator's Situation Awareness and takeover performance in a conditionally automated system with machine learning methods.

The purpose of this topic is predicting operators' quantified SA and takeover performance using their eye tracking data. Two types of eye tracking features were included: eye movement features related to fixations and saccades, and AOI features. Machine learning models for SA and takeover performance were developed using extracted eye tracking features as input. The analysis is conducted in both domains:

- **Drive study:** Using eye tracking features to predict drivers' Situation Awareness and takeover performance in a semi-autonomous driving system with machine learning methods.
- **Piloting study:** Using eye tracking features to predict pilots' Situation Awareness and takeover performance in an autopilot-assisted instrument landing with machine learning methods.

Research question: Can operators' SA levels and takeover performance levels be accurately predicted from their eye tracking features using machine learning models?

Hypothesis: It is hypothesized that Machine Learning classifiers trained on eye-movement features will achieve statistically significant predictive accuracy for classifying higher versus the lower SA group, as well as between those with higher versus lower performance group.

- H_0: Eye-tracking features cannot predict SA levels or takeover performance levels better than chance (e.g., accuracy not significantly higher than majority-class baseline).
- H_a: Eye-tracking features can predict SA levels and takeover performance levels better than chance, achieving meaningful classification performance.

The application of the outcome could be evaluating and monitoring driver's SA and performance levels automatically without human input in the future.

1.3 Research Contribution

Here are the research gaps identified in the three topics above and the contributions of the current thesis that address these gaps.

The current thesis focuses on operators' takeover process from autonomous control to manual control. In the driving domain, takeover process and TOR have been widely explored, including systematic review of TOR design [34], relevant factors that influence takeover performance [162] and eye tracking as a physiological approach to measure drivers' SA and takeover performance [172]. In the contrary, there is much less work that directly analyzes pilot performance during automation-to-manual transition process under malfunctions. Related aviation literature emphasizes the "out-of-the-loop" problem and autopilot surprise during unexpected events [60], as well as concerns about the loss of manual flying skills with frequent automation use [64]. This thesis addresses this gap by conducting a takeover experiment in the piloting domain, focusing on failure-triggered automation-to-manual transitions using the Instrument Landing System (ILS). It applies comparable research questions and methodologies as in the driving domain driving domain are applied, thereby providing new empirical data and insights into pilots' takeover behavior and its implications for safety and training.

When using eye tracking technique to study the operators' SA and takeover performance, prior work has proposed a number of eye tracking metrics to reflect operators' eye movement. One category includes non-voluntary eye movements such as fixation duration and counts, saccades [158] and pupil sizes [100]. Another category contains AOIs-based metrics, such as percentage of time and number of gazes inside pre-defined AOIs [80], [70]. However, most research uses only a limited subset of these features rather than analyzing a broad spectrum of eye tracking metrics. In the piloting domain in particular, definitions of AOIs vary depending on scenarios. Many studies also do not fully separate individual gauges on the instrument panel [62, 98], which may mask important differences in how pilots allocate attention among gauges during takeover conditions. This thesis addresses this gap by incorporating a comprehensive set of eye tracking features into the analysis, including metrics from all pre-defined AOI. Additionally, the AOI definitions involve the operator's entire field of view, enabling the exploration of not only central areas but also peripheral regions.

Additionally, the calculation of these metrics involves time-averaging operations. Previous research mainly focusing on one time phrase and investigate the eye movement metrics within this time period[94, 70, 77], therefore lacking comparison of eye movement between different time phrases. A limited number of studies did analysis on eye movement metrics on different time periods. Deniel *et. al.* [35] explored the takeover process by segmenting it into three temporal phrases, automation activated, takeover preparation and effective resumption of manual control. They explored how gaze behavior (e.g. dwelling time in different AOIs, average visit durations, possibly fixations) in these segmentation differs under different levels of agentivity (i.e. more vs less driver engagement). In the piloting domian, Lu *et. al.* examined pilots' fixation number and dwell time on two selected instrument panels during an ILS approach [98]. One of their main findings is that the less experience pilots fixated significantly less than more experienced group after the autopilot is turned off. These findings was limited to differences between groups across two time phases, but did not analyze how eye movement patterns changed within a single group across those phases. Overall, while these studies demonstrate that temporal segmentation can reveal meaningful differences, most existing work does not capture the dynamic changing of gaze behavior throughout the entire takeover process. For example, how operators' eye movement changes along time under different tasks context. This thesis addresses this gap by incorporating event-based TW segmentation into the analysis. All AOI-based eye-tracking features were examined within each TW, enabling comparisons not only between different groups but also across time within the same group. This approach highlights the importance of considering the task context when interpreting optimal eye movement patterns.

At last, although takeover from automation has been extensively studied in the driving domain, and the "out-of-the-loop" problem has long been recognized in aviation, direct comparisons between the two domains remain scarce. Most cross-domain works are conceptual or review-based, such as discussing shared "out-of-the-loop" problem [60, 106], or drawing lessons from aviation automation to inform driving automation design [15]. However, very few studies have measured takeover behavior in both domains using the same experimental logic or eye-tracking metrics. The design and methodology of the current thesis enable a direct synthesis of results across domains, offering a rare empirical basis for cross-domain comparison.

In all, the current thesis addresses these gaps by conducting parallel studies in both driving and piloting contexts using comparable takeover scenarios, identical eye-tracking metrics, and similar event-based TW segmentation. Together, the three topics lead to the following key contributions:

First, this thesis identified distinct eye movement patterns among operators with higher SA, better takeover performance, or belonging to specific population groups. This was achieved by comparing AOI-based eye-tracking features across multiple TWs, defined based on key events within the scenarios.

Second, the thesis highlights the importance of task context and time period when examining operators' eye movement. Different goals of each time window lead to different optimal eye movement patterns.

At last, the thesis deepens the understanding of the takeover process in two domains. Differences in control and monitoring demands between driving and piloting result in distinct eye movement patterns.

1.4 Thesis Structure

An overview of eleven chapters of the current thesis is provided below.

In Chapter 2, Chapter 2.1 introduces the takeover process in different conditionally automated systems, including autonomous vehicles in the driving domain (Chapter 2.1.1) and the ILS in the piloting domain (Chapter 2.1.2). The background of SA is introduced in Chapter 2.2 and the introduction of eye tracking techniques is introduced in Chapter 2.3.

In Chapter 3, the details about the driving and piloting experiments are presented. The experiment design, instruments, participants, and procedures of the driving and piloting experiments are elaborated in Chapter 3.1 and Chapter 3.2, respectively.

In Chapter 4, the details of data analysis pipeline is presented. Firstly, the data preprocessing procedures for the eye tracking data are introduced in Chapter 4.1. Chapter 4.2 then presents the SA and takeover performance metrics, followed by Chapter 4.3, which describes the TW separation for both domains. Finally, Chapter 4.4 details the data processing methods used in the current thesis.

Chapters 5–10 present six individual studies addressing three research topics in the driving and piloting domains, following a standard journal article structure. Chapters 5 and 6 report the driving and piloting studies for Topic 1, respectively; Chapters 7 and 8 present the driving and piloting studies for Topic 2; and Chapters 9 and 10 cover the driving and piloting studies for Topic 3.

Chapter 11 provides a final summary of the overall synthesis by offering comparisons across topics and domains, thereby enhancing the understanding of different eye-tracking features, and how eye movement patterns may vary across time windows with different task demands. Afterwards, this chapter addresses the limitations of the current thesis and outlines directions for future research. Finally, it discusses the applications and implications of the thesis findings.

Chapter 2

Background

Automation has spread into every section of life today with the development of intelligent computers and the necessity of reducing human labor. Regardless of the application area, automation can reduce human users' mental and physical workload [165]. However, it can also decrease their SA level if the operator is not familiar with the automation functions or is too reliant on it.

This chapter reviews previous research on transitions from automated to manual control in both driving and aviation domains. It begins by outlining the current state of automated control systems and their interactions with human operators. This is followed by an overview of the key factors influencing takeover performance in driving contexts, as well as potential challenges that may arise during the takeover process. Related topics in aviation, including automated control systems and real-world aircraft accidents, are also discussed. The chapter then introduces the concept of SA and summarizes existing methods used to assess it, highlighting its significance in the takeover process. Lastly, the review summarizes the current state of eye-tracking technology and explores its application in assessing SA.

2.1 Takeover process in different conditionally automated systems

2.1.1 Takeover process in driving

2.1.1.1 Autonomous vehicles development and taxonomy

The inception of autonomous cars can be traced back to the advent of cruise control, an electronic system that allows drivers to set their vehicles at a certain speed. The first car with a cruise control function can be traced back to 1965, which was introduced by American Motors. Subsequent to that milestone, human drivers progressively handed over their driving responsibilities to automated machines.

Building upon decades of experience with lower-level driving automation systems, such as cruise control or lane-keeping assistance, the automotive industry has recently achieved notable progress in higher-level vehicle automation. One prominent example is Waymo, which offers fully autonomous, rider-only services in multiple U.S. cities. This service is available to the general public through its app [160]. Other companies, such as Cruise, also pursued Level 4 automation; however, their operations were suspended following a safety incident and subsequent regulatory actions [88].

Shifting to the consumer market, Tesla remains the most prominent manufacturer offering vehicles equipped with autonomous driving capabilities. However, its Full Self-Driving (FSD) system continues to be recognized only as a Level 2 system, requiring full driver attention and intervention when needed [116]. Meanwhile, many other manufacturers are diligently striving to build commercialized autonomous vehicles. Mercedes-Benz has actually overtaken Tesla in certified Level 3 deployment. Their Drive Pilot system was certified for Level 3 operation in Germany in 2022 [7] and in some U.S. states (California and Nevada) in 2023 [107]. BMW's IX has introduced hands-free Level 2+ systems (like Highway Assistant in 2023 for U.S. and Germany), but not true Level 3 yet [19]. Hyundai advanced their Highway Driving Pilot to Level 3 functions, expected on the Genesis G90 in Korea and Europe [74]. However, considering the technology barrier and the regulation vacancy, commercializing self-driving vehicles with driving automation equals to or higher than Level 3 remains a challenge in the contemporary landscape.

From the inception of cruise control functionality to the evolution of fully autonomous vehicles, individuals have been exposed to a proliferation of vehicular terminologies, such as driverless cars, self-driving cars, or autonomous cars. A clear definition of the different levels of vehicle automation is essential for both industry and academia. It enables

accurate communication regarding the capabilities of autonomous systems across various manufacturers and research contexts. More importantly, it highlights the progression between automation levels, marking critical milestones in the advancement of automated driving technologies.

In 2014, Society of Automotive Engineers (SAE) first introduced the J3016 standard (SAE J3016™ Recommended Practice: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles) [135] to elucidate the different levels of driving automation. This standard has evolved into a preeminent reference within the contemporary landscape of driving automation research. In this standard, driving automation is divided into six discrete levels wherein Level 0 signifies no driving automation and Level 5 conveys fully automated driving. With detailed terms and definitions provided herein, the standard is widely used to explain how human drivers and control systems share different degrees of control roles in driving tasks. The standard's most recent iteration in 2021 has provided further elucidation about the border between Levels 3 and 4 of automation [75]. This revision serves to augment the precision in characterizing automation levels, as shown in Figure 2.1.

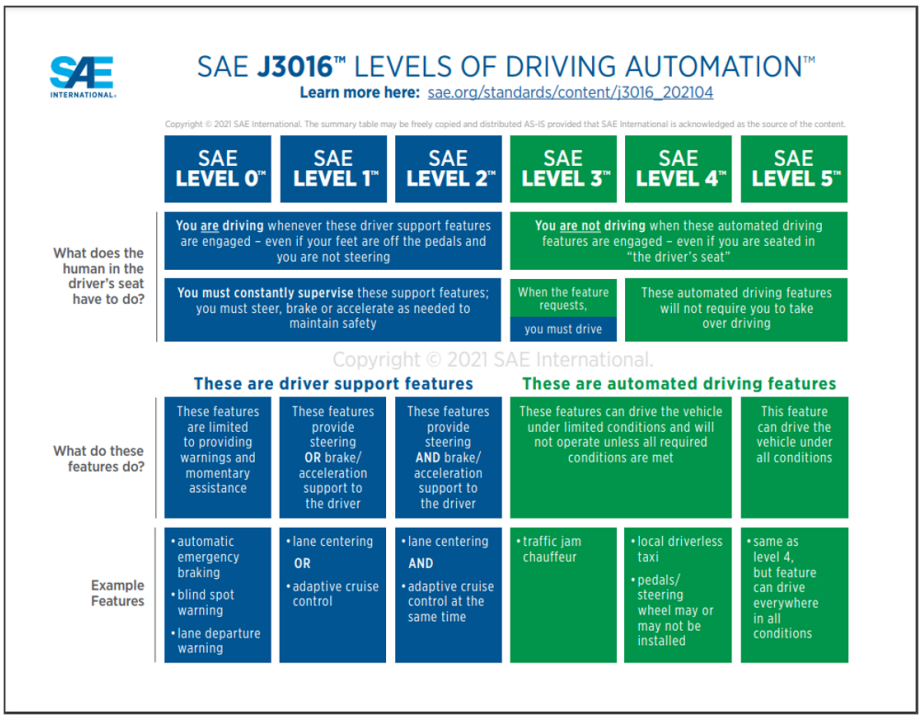


Figure 2.1: J3016 driving automation taxonomy [75]

The Level 3 driving automation is named "Conditional Driving Automation", where the entire Dynamic Driving Task (DDT) is performed by an Automated Driving System (ADS) within its Operational Design Domain (ODD), with the expectation that the user is receptive to ADS-issued requests to intervene [75]. In other words, a conditional automated vehicle is capable of performing transportation to people or goods between two spots without human intervention within its ODD but also expects a human driver to take over the DDT following ADS's request under system failures. It is essential to acknowledge that the term partially automated vehicle denotes a different concept. According to the SAE standard, the "Partially Driving Automation" corresponds to Level 2 of driving automation, where the automation system can control the lateral and longitudinal motion, but human drivers are expected to supervise the driving system and surrounding events during the whole time. Level 4 driving automation is named "Highly Driving Automation", which is similar to Level 3, but the ODD-specific automation system performs the entire DDT without any expectation that a user will need to intervene. This level requires automation to process a higher ability to manage unexpected events or hazards.

Nowadays, most vehicles on the consumer market are still between the Levels of 2 and 3, where the interactions between drivers and the autonomous system remain complicated, and the takeover from autonomous control could still be challenging.

2.1.1.2 Influential factors in Level 3 autonomous vehicle takeover

The most distinctive aspect of a Level 3 autonomous driving system is that it requires the driver to intervene in the driving task when requested, commonly referred to as the takeover process. This process includes the human driver disengaging the autonomous control and assuming manual control of the vehicle using the steering wheel and pedals. Compared to driving Level 2 vehicles, the drivers of Level 3 autonomous vehicles don't need to monitor the driving task, so they have the opportunity to engage in Non-driving Related Tasks (NDRTs), such as reading or texting [113]. As mentioned before, the main difference between Level 3 and Level 4 is the anticipation of driver intervention, so Level 3 drivers also need to be ready to takeover the driving task when they are requested to.

In a Level 3 autonomous car, the takeover process begins with a TOR, which is usually in visual, audio, or haptic modality to remind the driver to act. Within the context of conditional driving automation, the driver may engage in various NDRTs while the vehicle operates in automated driving mode, thereby relinquishing direct supervision of the driving task. The TOR operates as an alert to redirect their attention back to the surrounding situation and manual driving task. The driving automation system issues the TOR owing to its recognition of an imminent encounter with situations that exceed its ODD in the

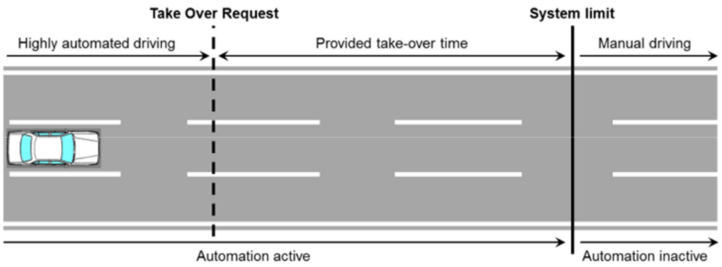


Figure 2.2: Takeover process [56]

near future. Consequently, it requires human drivers to take over the driving task before the system limit is reached. Following the perception of the TOR, drivers are expected to transfer into manual driving of the vehicle immediately. A visual representation of this process is provided in the subsequent Figure 2.2.

A substantial amount of research work has been dedicated to investigating the takeover process in Level 3 conditional autonomous vehicles. Collectively, these research works aim at studying different factors that affect the drivers' takeover process and intend to explore the possibility of improving drivers' performance. Factors that could influence the takeover process have been categorized into two distinct types: objective and subjective. The objective factors pertain to the inherent properties of the driving situation or automation system itself, whereas subjective factors are associated with the drivers' individual characteristics and adaptation to the conditional automation system [109].

As for the objective factors, time budget is one of the most important factors that influence the takeover behavior. Given the drivers might be engaged in other NDRTs during the period of highly automated driving, they need enough time to redirect their attention back to the vehicle upon the perception of a TOR. Gold *et al.* have conducted a comparative study of drivers' behaviors across 5 seconds TOR-time and 7 seconds TOR-time [56]. The result reveals that drivers become more likely to employ accelerator and brake inputs in situations with abbreviated decision-making intervals, indicative of shorter TOR times. Additionally, Wandtner *et al.* also tested drivers' performance under varying TOT budgets, finding that drivers tended to apply greater braking force in more time-critical situations [157]. Nevertheless, the TOT budget depends not only on the system design. If the vehicle is about to exceed its ODD by exiting a highway, the system can accurately anticipate the time point at which it will reach its operational limits and subsequently provide the TOR at an appropriate time. However, in situations wherein the ODD is exceeded by encountering unforeseen road hazards, it is uncertain how much time the system will leave the human drivers to takeover [161].

Also, the road condition has a great impact on the takeover process. Drivers are required to maintain varying levels of attention and collect distinct information when driving on different road types to avoid hazards and accidents. It has been well acknowledged that curved roads contribute to more crashes compared to straight roads, with the sharper bends exacerbating the situation [112]. Anderson, Bauer, Harwood, and Fitzpatrick found that the average radius of curvature, the average rate of vertical curvature, and predicted speed reduction are significantly related to accident frequency [3]. Researchers in New Zealand replicated a similar study undertaken in the United Kingdom, revealing a pronounced association between road curvature and accident rates [65]. Drivers may underestimate the curve radius and enter the curve at an inappropriate moment, thereby leading to an accident [21]. Curved roads impose higher demands on manual driving tasks since drivers need to collect more information to maneuver the car properly, and this trend also exists in the takeover process. Drivers exhibit higher TOT and lateral deviation when facing curved road situations in the takeover process [114].

TOR serves as an alert to remind drivers to redirect their attention back to the driving task. As the initial phase of this intricate human-vehicle interaction procedure, an effective interface would be crucial for a successful takeover [142, 14]. It could be deployed through visual, auditory, or haptic modalities which exert a significant impact on takeover behaviors. Visual modality can convey more complex information, such as speed, RPM, and fuel level, and it is easier to interpret. Auditory and haptic alerts can be detected without the driver's eye movement. However, it requires more effort to program the auditory and haptic information into a more usable pattern [73], so they can be easily understood and memorized by drivers in a limited amount of time [105]. A visual reminder could be an icon with pertinent information on the front windshield using a heads-up display. An auditory reminder could be a beeping sound or a verbal message from the vehicle's speaker [126]. The haptic reminder could be employed on various locations, such as the steering wheel, pedals, and driving seat [73]. Much research work underscored the effectiveness of multimodal TOR systems in enhancing takeover performance. Huang *et al.* compared the automated vehicle takeover performance under unimodal, bimodal, and trimodal combinations of visual, auditory, and tactile signals. Their findings revealed that multimodal signals with a tactile component facilitated faster takeover and improved information processing under conditionally automated driving [72]. Additionally, the fact that different NDRTs that draw upon the same perceptual resources as the takeover reminder highlights the insufficiency of relying on a single reminder modality. For instance, reading while receiving a visual reminder or listening to music while receiving an auditory reminder can cause interference. Therefore, it will be more effective if the reminder can combine with tactile signals [169].

In partially automated vehicles, drivers are expected to monitor the traffic condition while engaging in some NDRTs simultaneously, so a dual-task paradigm is assumed here. Nevertheless, one big improvement from Level 2 to Level 3 automation is that drivers are relieved from monitoring the driving environment, which means they don't need to decide whether they need to takeover based on the current driving condition. In Level 3 automation, the takeover requirement is initiated by the automation system rather than the human drivers, so drivers can focus on the NDRTs until they perceive the TOR. This change necessitates significant enhancement of drivers' attention on NDRTs, so there are more NDRTs that could be conducted in Conditional Automation Vehicles [113]. For everyday takeover tasks, drivers can interact with in-vehicle information systems, such as navigation, AC, or radio [149, 148]. They can also do reading, eating, or communicating with other passengers [24, 95, 170]. Nevertheless, in experiments, researchers normally use standardized NDRTs because they attempt to maintain the workload standardized [127]. Common standardized NDRTs could be tracking tasks, such as Critical Tracking Task (CTT) [76], Surrogate Reference Task (SuRT) [87], and n-back [81]. Different NDRTs have different modalities and cognitive workloads, so they affect drivers' takeover behaviors in different manners [127].

Compared to the objective factors, subjective factors influence takeover behaviors from the driver's perspective, such as drivers' experience, age, and cognitive ability. Among all of the subjective factors, driving experience is a critical determinant of takeover behavior and has been explored by researchers from multiple perspectives. Zhang *et al.* conducted an experiment to investigate the impact of driving experience (defined by licensed years and annual kilometers) on takeover performance in an automated system, and their findings indicated that drivers' previous driving experience is highly correlated with their takeover performance [170]. However, only using licensed years and annual kilometers to describe driving experiences is not adequate. The driving experience could also be referred to as either the experience with ordinary vehicles or the experience with automated vehicles. Hergeth *et al.* divided drivers into four different prior familiarization groups (no familiarization, description, experience, description, and experience) and reported their takeover performance in simulated driving experiments. The result shows that prior familiarization with TORs had a positive effect on their performance [67]. The experience group in this work gained their experience with TORs from the training session provided by the experimenter rather than in real life. This is reasonable because it can be quite challenging to recruit drivers with experience with TORs in their real life.

While driving experience positively influences the driver's takeover performance, it also has a close relationship with age, which sometimes causes other effects. Human cognitive ability, eyesight, and decision-making ability could decline as age increases. With a rapidly

aging population and a growing emphasis on independence, a growing body of research work is focusing on older drivers. Ding *et al.* conducted a level-3 autonomous driving simulation experiment involving drivers from three age groups. The results showed that older drivers had significantly higher HPT and statistically equivalent SAGAT scores compared to the other younger and middle-aged groups [40]. Additionally, the older drivers have significantly higher takeover time and lower takeover quality compared to the younger and middle-aged groups [37]. Anstey *et al.* reviewed the literature on relevant factors with safe driving behaviors of older drivers and found that cognition, sensory function, and physical function could be appropriate predictors of their driving abilities [4]. Horswill *et al.* found that older drivers need more time to perceive hazards and this result could be accounted for less contrast sensitivity and useful field of view [69]. Both cognitive ability and eyesight are critical to the safety and efficiency of the takeover process because the takeover process includes much information perception and comprehension. Thus, researchers explored the influence of age on drivers' takeover process under different driving conditions. Huang and Pitts found that older drivers with higher exercise frequency demonstrated shorter decision-making time in the takeover process of a semi-autonomous vehicle. Also, they stated that the benefit of exercise is more predominant to older drivers [71]. This conclusion verifies that drivers' cognitive ability impacts the takeover process through age. Older drivers also have different preferences on the NDRTs. Clark *et al.* compared the younger and older drivers' engagement of pre-designed NDRTs and their vehicle control performances in the takeover scenarios. They concluded that older drivers preferred electronic devices less than younger drivers, and they were more likely to brake harder after the takeover if they were more engaged in NDRTs [26].

Those factors are related to each other. For example, senior age usually indicates longer driving experience but lower cognitive ability and probably even lower trust in autonomous systems due to their strong manual driving habits. These correlations make the analysis of each individual factor challenging, so researchers usually control other variables when conducting the analysis. In all, the driver's takeover process is affected by both objective and subjective factors reviewed above. By studying the relationship between takeover results and influential factors, we could improve the takeover process by improving these relevant factors. The objective factors are usually related to the semi-autonomous driving system settings, encompassing TOR time budget, alert modality, and NDRTs, so the conclusions can be used as a reformation of the system. The subjective factors are mostly relevant to drivers' demographics or abilities, so they could be used in the driver training design.

Takeover is a complicated interaction process that requires human drivers to react timely, understand necessary information correctly, and conduct the subsequent manual driving safely. Similar metrics were employed in the studies reviewed above to assess

drivers' takeover performance, such as TOT, lane deviation, and brake force. However, the relationship between these performance metrics and drivers' eye movement remains underexplored, and individual differences in eye movement patterns also require further investigation.

2.1.2 Takeover process in piloting

2.1.2.1 Autopilot takeover process

Aviation stands as one of the domains in which automation has been implemented earliest and broadest. Common aviation automation systems include the flight navigation management system, aircraft pressurization system, automated fuel management system, auto-throttles and digital engine controls, and avionics flight plan system [2]. Besides the above avionics systems, Automated Flight Control System (AFCS), is one of the vital systems in modern aviation, aiding pilots in controlling an aircraft. Manually flying an aircraft is a closed-loop feedback process where the human operator needs to progressively apply a series of inputs to ensure the control purpose can be achieved smoothly(e.g., altitude change). Some maneuvers are exhausting or challenging for human pilots because they either require high accuracy control or long-time monitoring [122]. Moreover, the lack of visual cues can be problematic for pilots' manual flying as well. With faster response speed and qualified handling, AFCS plays an essential role in assisting the human pilot to overcome those challenges. It can provide a better flying quality, improve the primary performance of the aircraft, and lessen pilots' workload at the same time [103]. In many research materials, the words autopilot and AFCS are used interchangeably [134]. AFCS also includes the automatic functions employed on the Unmanned Aerial Vehicle (UAV) in some works. Autopilot systems are usually used in the range of manned aircraft, controlling the aircraft without the pilot directly operating. It helps to reduce human pilots' fatigue and operation errors during long-term flights by conducting the most time-intensive and non-decision-making tasks [78]. In the current work, we use the word "autopilot" to refer to the automated system that performs a series of maneuvers on a manned aircraft.

The autopilot function is irreplaceable in increasing the aircraft's operational capability and reducing pilots' workload. Still, human pilots are always required in the current general aviation and commercial airlines. Pilots are required to keep monitoring the automation system, which is responsible for lower-level tasks, such as information acquisition, information analysis, and action implementation [124]. When the autopilot function malfunctions, pilots must quickly detect the issue and take appropriate actions, such as rebooting the system or switching to manual control with minimal delay. An operator's lack of necessary

knowledge of the autopilot system and flight status can lead to tragic accidents. China Airlines flight 140 from Taipei to Nagoya airport in Nagoya, Japan, in April 1994, is an example [118]. During the approaching phase, the co-pilot unintentionally selected the takeoff/go-around (TO/GA) mode, which trimmed the aircraft nose up, and the pilots could not return the TO/GA off due to the training mismatching with the current plane autopilot settings. The pilots kept attempting to fly the airplane nose down with the autopilot function working simultaneously, which eventually led to a crash. A similar tragedy happened in 2013 with Asiana Airlines flight 214 from Incheon International Airport to San Francisco [130]. The flight was attempting to conduct a visual approach, but the plane was too slow and too low due to the captain selecting an inappropriate autopilot mode, which led to the autothrottle failing to maintain its speed. The pilots detected the stall situation too late, and an impact happened at last. Unlike the driving domain, the transferring procedure from autopilot control to manual flying does not have a commonly used terminology. In the current proposal, we keep the term "takeover" to refer to the transition period where an automated system controls the plane at the beginning and changes into manually flying afterward.

Many researchers have investigated pilots' behaviors when facing autopilot failures and tried to identify factors influencing the autopilot takeover result. Overall, familiarity with the autopilot system settings is one of the critical factors in a successful takeover. Beringer and Harris Jr. have conducted experiments to explore pilots' responses when facing autopilot malfunctions. The result showed that pilots used significantly less time to disconnect autopilot if no other significant actions were taken after the malfunction was detected, compared to those who manually overridden the autopilot first. Also, a significant percentage of maladaptive responses led to severe accidents [18]. Sarter *et al.* asked the pilots to fly a one-hour session with challenging autopilot events on a flight simulation, such as experimenter-induced mode transitions and revision of cruise altitude, with eye tracking data collected. And their eye movement revealed that they did not understand the mode (autopilot/manual) annunciation sufficiently [138]. Dehais *et al.* used eye tracking techniques to study pilots' behaviors when facing the "automation surprise", which refers to the situation when pilots are surprised by the behavior of the autopilot. Most pilots failed to deal with the situation in an acceptable amount of time and also exhibited shorter fixations and saccades, which indicates they were actively searching for information [31]. The interaction mode between autopilot and human pilot also has a dominant effect on failure detection. Ephrath and Curry employed four different levels of participation in a simulated low-visibility landing approach to explore the impacting factors of failure detection. Monitoring, lateral axis manual, longitudinal axis manual, and fully manual were included in the experiment, and the result showed that detection times in a manually controlled axis

are significantly longer than detection times in an automatically controlled axis [52].

A successful takeover in the piloting domain involves timely detection of autopilot malfunction and safe manual control afterward. Although much research has explored the process from autopilot to manual takeover transfer, little work has utilized the concept of SA in this process. The current thesis will use the concept of SA to investigate the piloting takeover process in the later content.

2.1.2.2 Instrument Landing System

The ILS is a typical example of an autopilot system which can be dated back to 1920s [137]. At present, ILS is one of the primary worldwide, International Civil Aviation Organization (ICAO)-approved, precision landing systems [150]. ILS is a runway approach aid system based on two radio beams that provide pilots with vertical and horizontal guidance during an approach to land. It comprises two subsystems, one on the ground and the other on the aircraft. The ground part includes the Localizer (LOC) and the Glideslope (GS). The antennas of LOC and GS are located on the end of the runway and the aerodrome, respectively, as shown in Figure 2.3.

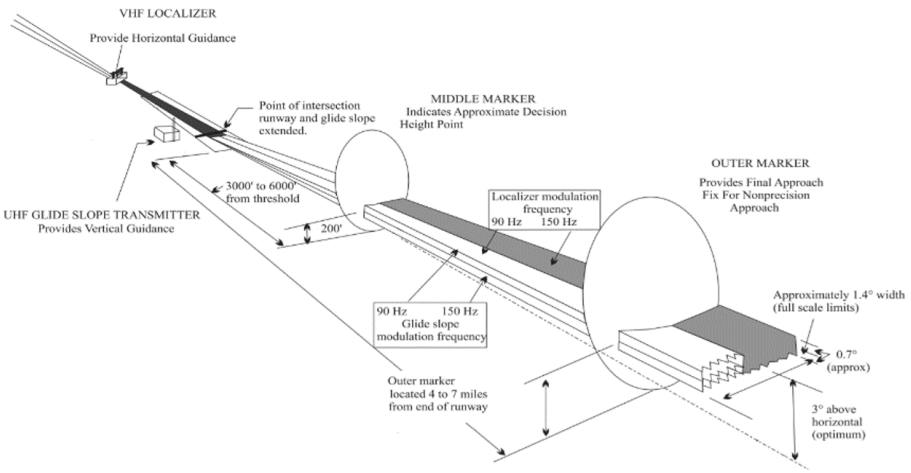


Figure 2.3: Instrument Landing System [13]

The LOC (108 to 112 MHz) provides horizontal guidance by transmitting two narrow beams intersecting at the runway centerline. Figure 2.4 shows the shape of the LOC beam viewed from the top of a runway. Similarly, the GS (329.15 to 335 MHz) provides vertical guidance with the other two beams, intersecting at the correct vertical descent profile,

which is three degrees in accordance with the standard of ICAO. Figure 2.5 shows the shape of its beams viewed from the side of a runway.

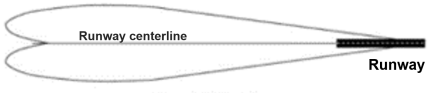


Figure 2.4: ILS Localizer

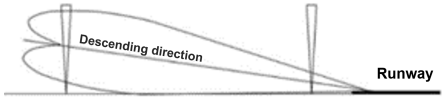


Figure 2.5: ILS Glideslope

Another subsystem of ILS located on the aircraft includes the beam receiving antennas and the ILS receiving screen on the cockpit, as shown in Figure 2.6. When landing with ILS, the pilots would first attempt to align with the runway centerline under the instruction from the Air Traffic Controller (ATC). To do so, pilots need to align the LOC pointer, which is the vertical line on the ILS receiving screen, in the center. Then, the pilots need to level off the aircraft until the GS pointer, which indicates the descending slope, is in the center of the screen. Once both the LOC and GS are intercepted, the aircraft would follow the fixed trajectory until it reaches the Decision Height (DH) point, where the pilot needs to have adequate visual reference to land. If not, the pilot must execute a go-around or be diverted to another airport. According to the standard of ICAO, different types of aircraft should fly different categories of ILS where different Runway visual ranges (RVRs) and DH apply [120].



Figure 2.6: ILS receiving screen

Like other autopilot systems, the ILS system relieves the human pilot from long-time maneuvering and monitoring. It also increases flying safety and efficiency by allowing aircraft to land in limited visibility conditions, such as at night or in bad weather. Landing with ILS is a common scenario in the experiment design of many previous aviation studies [52, 93]. In the current proposal, ILS is also used in the piloting experiment.

2.2 Situation Awareness

The concept of SA was first proposed by Endsley in 1988, which is briefly defined as the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future [45]. This concept was initially proposed when human factors researchers were studying pilots' internal models of the world around them and trying to improve their mission effectiveness. As more and more information was presented to pilots through sensor systems, weapon systems, and control systems, it became increasingly challenging for pilots to successfully perceive and understand all this information with their limited cognitive abilities. Thus, the SA concept was defined in the pilot-machine interaction process, as depicted in Figure 2.7. SA is the initial input in the process, which provides an important foundation for the subsequential actions. In dynamic and complex systems, operators must continuously perceive and process information from multiple sources. How they allocate their limited attentional resources significantly affects their SA. Over-focusing on specific elements can lead to critical oversights and potentially fatal consequences [84].

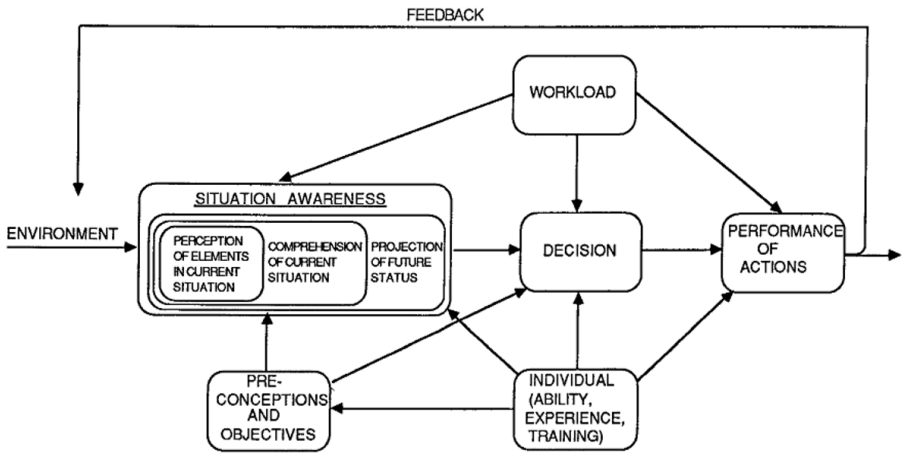


Figure 2.7: Aircrew decision model [45]

Endsley further elucidated the SA into three ascending levels: the perception, the comprehension, and the projection of information [45, 44]. In different fields, the elements that require attention are different. For example, when driving a vehicle, level 1 SA needs drivers to know the elements like location, direction, destination, etc. Level 2 SA requires the driver to comprehend the meaning of this information—for example, understanding fuel levels or time constraints. Level 3 SA demands the drivers to project that information

to the future trajectories and operations. When operating an aircraft, the level 1 SA needs pilots to know the elements like location, altitude, heading, etc. The level 2 SA requires the pilots to comprehend the remaining resources, the mission, and the strategy. Level 3 SA demands the pilots to project the aircraft's tactics and maneuvers. Factors like geographical awareness are needed across all three levels [47].

SA is an important measurement when researchers try to understand the operators' performance in automation systems. Automation can improve system reliability and performance by reducing human workload human errors [117]. However, implementing automation in complex systems can cause troubles as well. Operators perceive information passively rather than actively during the automation period, which could delay detecting system errors or neglect the need for manual takeover [28]. Endsley and Kiris used the term "out-of-the-loop performance" to describe this circumstance and stated that it mainly resulted from the loss of SA. They conducted a decision-making experiment to prove that the negative aspect of automation could be avoided by determining the best level of automation. The result showed that participants had poorer SA in fully automated condition than when they had been in the manual condition. The intermediate Levels of automated condition (such as Level 2 among five levels) had an intermediate level of SA. Considering other advantages of automation, the out-of-the-loop performance might be avoided by implementing a partially automated system rather than a fully automated system [49].

2.2.1 Situation Awareness in driving and piloting research

Previous research regarding human interaction with automation systems often aims at improving SA level, which could lead to better performance. Those works mainly focused on understanding operators' information processing, improving their performance, and giving implications for future design.

In Level 3 automated driving, the control system handles the driving tasks while the drivers are engaged in the NDRTs, so they may be unclear about the current driving status when they are requested to takeover the car. A great amount of research focused on solving this by improving the SA. Research has shown that higher SA levels can result in a better ability to retrieve the control [154]. White *et al.* compared the "top-down" guidance and "bottom-up" guidance of TOR by testing whether they can help the drivers to rebuild their SA [164]. Tan and Zhang conducted a web-based experiment to figure out the effect of the TOR lead time in non-critical takeover situations on drivers' SA. The result showed that 25 to 30 seconds was the best for their SA, which could be used as implications for future TOR design [145]. Meanwhile, in pilot training, the ability to gain

enough SA is an important nontechnical competency for avoiding accidents [46]. Human errors related to SA contribute to 88% of accidents among major airlines [44]. Endsley mentioned that a training program specifically focused on improving SA would be more beneficial to pilots after they master basic flight skills. For example, pilots will be trained to construct Level 2 and Level 3 SA from the basic information available [48]. Muehlethaler and Knecht proposed a simulation-based SA training program with two theoretical parts and two identical fight sessions. They used a normal scanning pattern in the first fight session, and improvement in their SA levels and strategies was provided to them afterward before the second flight session [111].

2.2.2 Situation Awareness measurements

SA can be measured via various approaches. According to the definition, SA refers to the perception, comprehension, and projection of the surrounding elements, so it can be directly measured by testing whether the operator has observed and understood relevant information, such as the current status of systems, the environment condition, and the surrounding objects [30]. One way to achieve this is to use freeze probe techniques where the task performed by the operator is randomly frozen and the operator is required to answer questions about the current environment based on their memory and knowledge [115]. SAGAT, proposed by Endsley in 1988, is a typical example of this. Endsley proposed to measure a pilot's SA in the following manner. When the pilot flied a given mission scenario in a man-in-the-loop simulation, the cockpit and out-the-window displays were blanked at some random point. The pilot was asked some questions relating to their SA requirements selected from the whole SA questions set. The comparison of the true situation and perceived situation was an objective measure of the pilot SA [51]. Although SAGAT was developed in the aviation domain, it can be applied in various human factors research areas. The research work of HOGG *et al.* gave a great example of adapting the SAGAT in the assessment of the nuclear power plant operator's SA. As an extension of SAGAT, the method Situation Awareness Control Room Inventory (SACRI) was proposed in their work and was tailored to the characteristics of process control, including altering the questions asked, the time points those questions presented, and the content of testing scenarios [68]. Van dam *et al.* also employed SAGAT in their driving study where drivers' SA were tested after they received audible text messages. The SAGAT probes happened at 10 seconds and 30 seconds after the message was delivered, and the result showed that their SA levels were negatively affected at least for 10 seconds but no longer than 30 seconds [152].

The freeze probe techniques can provide a timely and objective assessment of operators'

SA, but it also brings its main disadvantage: intruding on the operator's task performance, which makes it hard to be used in real-world situations or time-sensitive scenarios [115]. Real time probe techniques can be used to eliminate this intrusion, such as Situation Present Assessment Method (SPAM) [43]. In SPAM, questions asked are similar to those in SAGAT, but the operators continuously conduct the tasks during the queries. Loft *et al.* used SPAM to examine participants' SA in simulated submarine track management. The time elapses between the queries and the answers are collected as indicators of their SA. The result showed that it took them a longer time to answer SPAM queries that assessed SA of the future display situation compared to the current display situation [96].

Furthermore, post trial techniques are the most easily implemented measurement. It can completely remove the intrusion of experimental tasks. The Situation Awareness Rating Technique (SART) is a post trial technique widely used by many researchers [146]. Saleem and Kleiner collected pilots' opinions on their own SA levels using SART scales after they performed landing approaches during different visual conditions for both Visual Flight Rules (VFR) and Instrumental Flight Rules (IFR) flights. The SART questions are divided into three aspects: demand, supply and understanding, and the total scores provide the final SA. The result showed that VFR-only pilots would have reduced SA compared to IFR-rated pilots [136]. Kohn *et al.* used SART as the measurement of drivers' SA when exploring the influence of interruption of non-driving tasks on their takeover performance during autonomous driving. The result showed that their SA levels were significantly different between scene interruption and baseline [86]. SART is highly correlated with the interviewee's confidence level, and it determines how an operator will act on that SA [50]. SAGAT method provides a more objective assessment of SA as a freeze probe technique, while SART is prone to be more subjective. Those different measurements all have their advantages and disadvantages, and they should be selected based on experiment conditions and research purposes.

Besides the direct measurements introduced above, SA can also be evaluated from a series of physiological approaches. The physiological data is collected by a series of physiological sensors, which provides an surrogate evaluation of SA. A great amount of research has tried to build the relationship between physiological data and operators' direct measured SA. Commonly used SA measurements and evaluations are summarised in Table 2.1.

Cardiovascular evaluations are widely used in previous SA studies. Key features such as Heart rate (HR), Mean HR and Heart rate variability (HRV) can be extracted from the raw cardiovascular data. HR is the number of heartbeats per minute. Mean HR is the mean of the operator's HR in each trial, and sometimes the value will be compared to the baseline. HRV is the fluctuation in the time intervals between adjacent heartbeats, which could

Table 2.1: SA measurements

SA direct measures	Freeze probe technique	SAGAT, SACRI	Pro: Objective and quantifiable; Con: Requires freezing the task.
	Time probe technique	SPAM	Pro: Real-time; Con: Interrupts operation.
	Post trial technique	SART	Pro: Non-intrusive; Con: Subject to self-report bias.
Physiological evaluations	Eye tracking	Fixation duration, Saccades angle, entropy, etc.	
	Cardiovascular	Mean HR, RMSSD, LF/HF	

be examined from time and frequency perspective [102]. From the time perspective, the metric RMSSD (root mean square of successive differences between normal heartbeats) is commonly used [140]. From the frequency perspective, the LF/HF ratio, which is calculated by the ratio of low frequency to high frequency in each trial, is a typical metric used by researchers. Generally speaking, increasing HR and LF/HF and decreased RMSSD value indicate increased workload fatigue and stress [147]. Some research work stated the correlations between SA and cardiovascular measures, though the strength of these correlations varies depending on the study. Wei *et al.* examined operator's SA while they were using different aircraft Cockpit Display Interfaces (CDI) and their HR were also collected in each scenario. The result shows significant differences between both SA and average HR among three different CDIs [163]. However, other studies did not discover statistically significant correlations between SA and cardiovascular measures. Mehta *et al.* examined the operator's SA and physiological load, using HR and HRV in different simulated offshore well control scenarios with and without failures presented. Although elevated heart rate was observed with drilling and failure scenarios, no significant difference between those two cardiovascular measures was found between different scenarios [104]. In all, the validity of the correlation between cardiovascular measures and SA levels is still uncertain, and different experimental environments might cast an influence on it.

Eye tracking is the most reliable and widely used among all physiological evaluations for SA [171]. It can provide numerical predictions of the specific level of SA as opposed to a mere statistical measure that may indicate cursory relevance with SA. The details will be discussed in Chapter below.

2.3 Eye tracking techniques

Research works using eye tracking techniques have gone through proliferation in the last 20 years (see Figure 2.8). This situation results from the increasing accessibility of high-accuracy eye trackers. Human eyes are vital for information gathering because they serve as the primary input device for visual information, which takes most of all information from the environment, especially when an individual interacts with a complex control system, such as a semi-autonomous vehicle or airplane. The perception of information is the first level of an operator's SA, and information gathering before that starts from human eyes. The operators are required to process a great amount of information while they are controlling a complex system, so "where they looked" has a great impact on their SA and task performance.

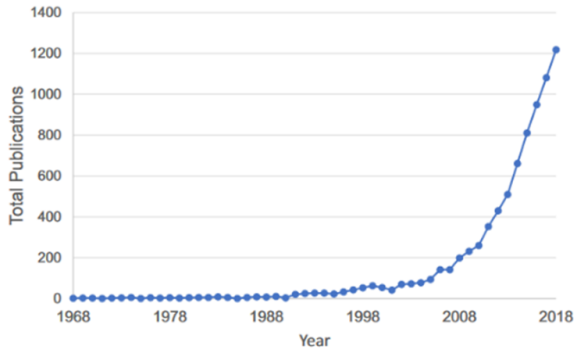


Figure 2.8: Number of publications using eye tracking techniques [24]

Individuals' eye movement can be tracked by eye tracking devices, also known as eye trackers. Eye trackers are mostly camera-based, but some eye trackers also use lasers to track the pupils. The view of the wearer can be tracked using the eye tracker's front camera. Eye trackers are also able to track unconscious eye movements resulting from involuntary body reactions to environmental stimuli, such as blink rate and pupil diameter [171]. Figure 2.9 shows a few pictures of commercial eye trackers available today

Saccade and fixation are two basic components in eye movement. Saccade is the period of time when the eyes are moving, and the fixation is the period of time when the eyes remain fairly still [128]. Krueger *et al.* gave a perfect visualization of participants' saccades and fixations in Figure 2.10, illustrating the temporal and spatial movement of eyes [83]. Previous research proved that information can only be visually gained while the eye fovea area, which is a small portion of the visual field, is pointing at a certain target [61].



Figure 2.9: Eye tracking glasses examples

Therefore, to keep gathering enough information through eyesight, individuals need to keep moving their eyes so the saccades and fixation will appear alternately. The saccades, fixations, and gaze points can be tracked and recorded by an eye tracker for further analysis. Those eye movement data reveal different information processing behaviors, and a series of eye tracking features could be extracted from them, which will be used in future analysis.

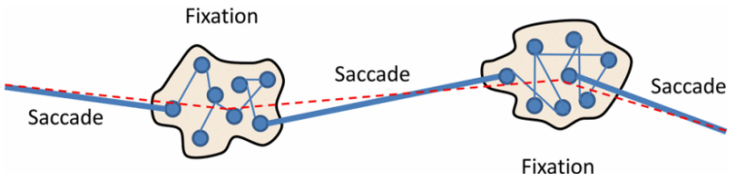


Figure 2.10: Fixations and Saccades [83]

The selection of eye-tracking features is guided by prior research, the specific objectives of the current study, and the characteristics of the experimental context. Eye tracking features calculated from saccades- and fixation-related values are fundamental features employed by many previous research [92, 20, 168], such as total fixation count, average fixation time length, standard deviation of fixation duration, etc. Meanwhile, AOIs, which indicates areas that contain important information, is also a frequently used concept in building eye tracking features. Within the domain of human-interface interaction, the AOIs are usually assigned based on the locations of instruments and displays on the control panel [32, 133]. For example, in a flight study, as shown in Figure 2.11, the AOIs used include speed indicator, attitude indicator, navigation display, heading, windows, etc. [90]. In previous driving studies, commonly used AOIs include ego vehicle mirrors, dashboard, road signs, surrounding vehicles, pedestrians, etc. [6, 36]. After the definition of AOIs, many features could be further created, such as dwelling time, which is the total time that an individual spends in those AOIs, and entry counts, the total count of entries of those AOIs.

Time series theories can also be applied to the analysis. Researchers have used the

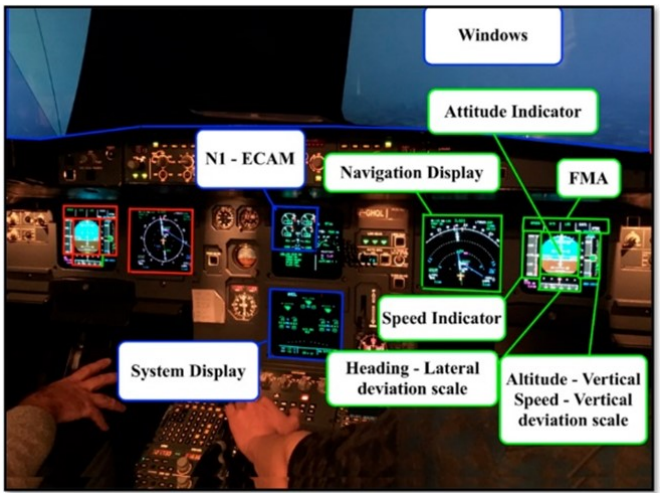


Figure 2.11: AOIs layout in a cockpit [90]

Markov chain model, N-gram, and entropy concepts to build eye tracking features [97]. Gaze entropy has gained more and more attention in recent years due to its ability to describe the average information or uncertainty associated with choices [141, 143]. It is a well-developed metric representing people's gaze behaviors quantitatively. It measures the variability and distribution of fixations, reflecting how a person's gaze is distributed across various AOIs. Ayala et al. investigate the impact a dual task on pilots' gaze behavior, which was represented by both traditional dwelling time and entropy-based analyses, including Stationary gaze entropy (SGE) and Gaze transition entropy (GTE) [8]. In this study, as well as in related work [12], SGE value was used to quantify fixation disperse, where a lower SGE value indicates concentrated fixation distribution among pre-defined AOIs and a higher SGE value indicates dispersed fixation distribution among these AOIs. Meanwhile, GTE depicts the sequential scanning behaviors, where a lower GTE value reflects a more predictable pattern and a higher GTE reflects a complex pattern of scanning. These results are also confirmed in other research work [9, 143].

Also, specially designed features were implemented as well [90]. Researchers have also specially designed models to calculate eye tracking features that reflect the sequences of pilots' eye gaze on various AOIs. The scan patterns for all trials may be recorded and the number of different patterns was as an eye tracking feature [90]. Similarly, Brams *et al.* recorded the sequences of AOIs in their study, then used the match ratio between the instructed gaze patterns and current gaze sequences as an eye tracking feature [20]. Table 2.2 below summarizes three categories of eye tracking features discussed above.

Table 2.2: Eye tracking features

Feature category		Examples
Fixation/Saccades-related feature		total fixation count, average fixation time length, standard deviation of fixation duration, average saccade duration.
AOIs-related feature	Basic	Dwelling time, Entry count
	Time series	Markov chain model, N-gram, entropy
Specially designed feature		The number of different combinations of AOIs, The ratio of the number of patterns matching instructed scan pattern over the total number of patterns.

2.3.1 Eye tracking and Situation Awareness

Eye tracking features can be applied in a variety of contexts. Research has shown that individuals' eye movements reflect their cognitive processes and task performance. Gaze patterns of people from different backgrounds can tell the diverse gaze patterns they implemented. They can also be used to train the machine learning model.

As previously discussed, eye tracking measures are the most reliable and most commonly used among all physiological evaluations for SA [132, 171, 153] in both driving and piloting domains. Many eye tracking features were positively correlated with operators' SA in multiple experimental scenarios. In the flight simulation experiment, van de Merwe *et al.* explored the applicability of eye tracking metrics as an SA assessment. The results showed that fixation rates and dwell time could be informative for the level 1 SA, and the use of entropy could be a Level 3 SA indicator [153]. In the semi-autonomous driving scenarios, Liang *et al.* discovered better overall SA correlates with longer time spent viewing the driving scene and more dispersed visual attention allocation. Also, location-based metrics, such as total fixation count, gaze count in AOIs, are best at differentiating between task conditions with significantly different SA [94]. Yang et al. also investigate how entropy is correlated with comprehension in SA for autonomous driving [167]. In this study, the SGE was employed to measure gaze distribution over a specified period. The more equally the fixation is distributed, the higher the SGE, indicating a searching gaze behavior.

Instead of testing the statistical correlation between SA and eye tracking features, Zhou *et al.* used the model Light Gradient Boosting Machine (LightGMB) to predict the SA scores with eye tracking features and got great accuracy [172]. Predicting SA with eye tracking features has a profound significance because of its nonintrusive continuous char-

acteristic [129]. It can avoid essential problems brought by traditional SA measurements, such as relying only on participants' memory and removing them from the ongoing task [29]. If the eye tracking features could be automatically used to predict the operators' SA, lively monitoring of their cognitive status would become possible.

2.3.2 Eye movement of older drivers

Previous research has shown that drivers from different age groups demonstrated different eye movement patterns.

In terms of driving-related visual patterns, previous research has consistently shown that older drivers tend to make more frequent, smaller saccades and exhibit a higher number of fixations. Maltz *et al.* conducted an experiment in which participants viewed photographs of traffic scenes from a driver's perspective. Their results revealed that older adults were characterized by a greater number of fixations and shorter saccades, while maintaining equal average fixation durations. Moreover, they devoted a larger proportion of their visual scan time to a limited subset of areas within the image [99]. Similar results were summarized by Carr *et al.* in a review paper, which reported that older drivers often display more frequent gaze fixations when moving into roundabouts and intersections. The review also highlighted the increased presence of "looked but did not see" (LBDNS) phenomenon among older drivers, whereby relevant hazards are fixated but not consciously processed [23]. Sun et al. further demonstrated that older drivers produced more frequent fixations when driving through roundabouts, whereas on straight road segments they tended to fixate on certain objects for longer durations [144]. These findings suggest that older drivers' increased fixation frequency and shorter saccades may represent a compensatory visual strategy to maintain SA under higher cognitive demands. Importantly, Sun et al. also reported significant associations between fixation patterns and cognitive measures (e.g. processing speed and divided attention), particularly during complex maneuvers like right turns [144]. These findings imply that older drivers tend to restrict their visual attention to a limited set of regions, especially under more complex tasks or in time-critical situations, rather than distributing attention broadly across the visual field [42, 85, 79]. Related research using driving simulators supports these results, showing that older drivers often adopt narrower visual scanning strategies, with longer fixations on the road ahead and fewer mirror checks, which may limit peripheral hazard detection [151].

Building on the understanding of how older drivers move their eyes in conventional driving scenarios, it is important to consider how these patterns may change during takeover process. Takeover process of the level-3 autonomous vehicle is a complex and time-sensitive

task that requires drivers to quickly switch to manual vehicle control. Especially drivers are allowed to do NDRTs, which means they are not necessarily monitoring the driving scenes during the autonomous control phrase. Therefore, drivers' visual attention and eye movement behavior becomes very important, as they reflect how well drivers are able to regain SA. As for older drivers, previous research has shown age-related declines in cognitive function, visual acuity, and decision-making ability.

In takeover scenarios, previous research have reported that older drivers exhibit significantly higher HPT, TOT and significantly lower takeover quality to the other younger and middle-aged groups, as reviewed in Section 2.1.1.2 above. Regarding for their eye movement patterns, similar results were observed as in regular driving scenarios. Specifically, older drivers exhibited longer fixation point duration and smaller saccades amplitude compared to younger drivers, indicating poorer visual exploration in a partially automated driving scenarios [123]. Likewise, older drivers were found to have longer gaze duration and smaller pupil areas in some takeover situations, reflecting their lower visual processing capacity [159]. In terms of gaze distribution and location, Molnar *et al.* conducted a simulated takeover experiment containing both manual and automated driving modes with younger and older drivers [108]. Their findings revealed that older drivers' gaze was narrower, especially in the vertical direction than younger drivers, but only during automated driving. Additionally, older drivers spent less time with eyes off road than younger drivers in both modes. This indicated that older drivers demonstrated a more "centralized" gaze orientation in the automated scenarios. Similar findings were reported by Peng *et al.*, who found that elder drivers focused significantly more on the road ahead than young group, and omitted checking side mirrors or rear mirror at the same time during the takeover process [125]. Additionally, they also showed significantly fewer transitions between different regions and are more influenced by high traffic density.

Together, these findings reveal a consistent pattern: older drivers tend to exhibit more frequent fixations, shorter saccades, and more centrally focused gaze strategies across both conventional driving and takeover scenarios. These eye movement patterns may be related to their age-related declines in cognitive and visual processing, and may limit their ability to rapidly regain full SA and vehicle control during the takeover process. This underscores the importance of considering age-specific visual patterns when designing support systems or interventions to facilitate safe and effective takeovers in level-3 autonomous vehicles.

Chapter 3

Experiments

In the process of transitioning control from an autonomous system to a human operator, perceiving and comprehending all relevant information in a limited time period poses a significant challenge for humans, which can lead to hazardous situations. For example, pilots may fail to respond appropriately during an autopilot malfunction if they are suddenly required to fly manually. This risk is exacerbated when pilots do not recognize their confusion about the autopilot's active mode, leaving them unprepared for the manual takeover and increasing the likelihood of an aircraft stall. Therefore, it is crucial to investigate the takeover process from autonomous systems to manual control.

To enhance the safety and efficiency of takeover processes in the domains of driving and piloting, one takeover driving experiment and one takeover piloting experiment are conducted in simulated environments. In both experiments, the "takeover" tasks were incorporated into corresponding scenarios, and operators' eye movements were recorded throughout the process. This thesis used an observational study design to explore eye movement differences between operators with different SA levels and takeover performance levels. No independent variable was manipulated. Instead, operators were grouped retrospectively based on their observed SA and takeover performance. This approach is commonly referred to as post-hoc grouping. As such, the analysis reflects naturally occurring differences in operators' behavior. Additionally, the thesis explored the eye movement between operators from different population groups as part of the observational study. The population groups (e.g. age, experience) were pre-existing groupings based on people's demographic characteristics rather than experimental manipulation.

Operators' eye movement were collected by eye tracking glasses throughout the experiments. By utilizing statistical and machine learning methods in the data analysis, we aimed

at exploring the eye movement from operators with different SA, as well as their takeover performance. The SA level and the performance are two critical, sequential components of the takeover process, during which operators first perceive and comprehend the information they gather and project it to a series of actions later, as illustrated in Figure 2.7 and discussed in Section 2.2. In the takeover process, the SA level reveals the operators' cognitive consciousness of the information from the environment, while the performance is related to any physical actions of the operators, like stepping on the brake or maneuvering the yoke. Meanwhile, the eye movements of operators from different population were also explored. These comparisons aimed to reveal how factors such as age or experience might influence visual behavior during the task.

The driving experiment is presented in Section 3.1, and the piloting experiment is introduced in Section 3.2.

3.1 The driving experiment

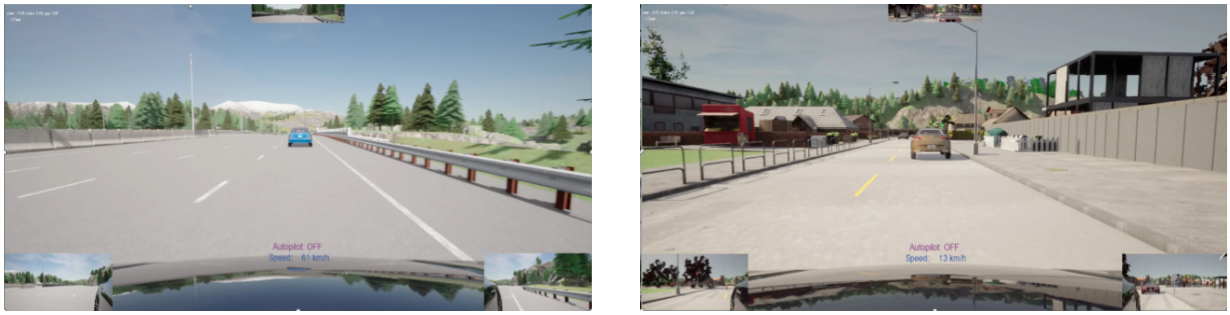
The experimental protocol was approved by the Office of Research Ethics of the University of Waterloo (ORE # 42299).

3.1.1 Driving experiment design

A series of simulated driving scenarios, representing trips from point A to point B, were developed for the current experiment. The takeover tasks in the current experiment were initiated by the automated control system at specific points during the scenarios, requiring the drivers to change into manual driving with a minimum time delay. These scenarios incorporated diverse driving conditions and takeover experiences to better reflect real-world situations and enhance the generalizability of the findings. The slight variations between trials also provided more naturalistic driving variability. Since each road condition required a different amount of time to complete, the duration of each trial varied slightly, preventing drivers from anticipating the upcoming events. Importantly, these variations were not designed as formal experimental manipulations but served to enhance the ecological validity of the simulation. Since the experiment used an exploratory observational design, the focus of the analysis was on post hoc grouping of participants based on their performance.

To provide diverse driving conditions, three different drive types were included: manual driving, automation without Non-driving Related Task (NDRT), and automation with NDRT. The latter two automation situations included takeover tasks, requiring drivers to

assume manual control when requested by the system. The drivers drove these different drive types on four different drive types: highway straight (HS) road, highway curved (HC) road, city straight (CS) road, and city curved (CC) road. Figure 3.1 shows the scenes of the highway and city road driving scenarios at the beginning. During the automation with NDRT scenarios, the NDRT, drivers engaged in NDRT to simulate common distractions, such as messaging or talking, that may occur during periods of automated control. Three NDRT were implemented on a separate laptop beside the driving seat, including two visual tasks: CTT [76] and SuRT [101], and one auditory task, N-back [81]. The CTT task required drivers to track a moving line and use up and down keys to keep the line aligned with a reference line in the center of the screen (see Figure 3.2a). The SuRT task asked drivers to find one unique circle amidst a sea of identical circles displayed on the screen (see Figure 3.2b). The N-back task asked drivers to recall numbers they heard in the sequence spoken by the experimenter, with the level of difficulty determined by the "N" steps back in the series (see Figure 3.2c). Drivers from different age groups were recruited to join the experiment, including young, middle-aged, and older drivers; detailed information will be covered in Section 3.1.3.

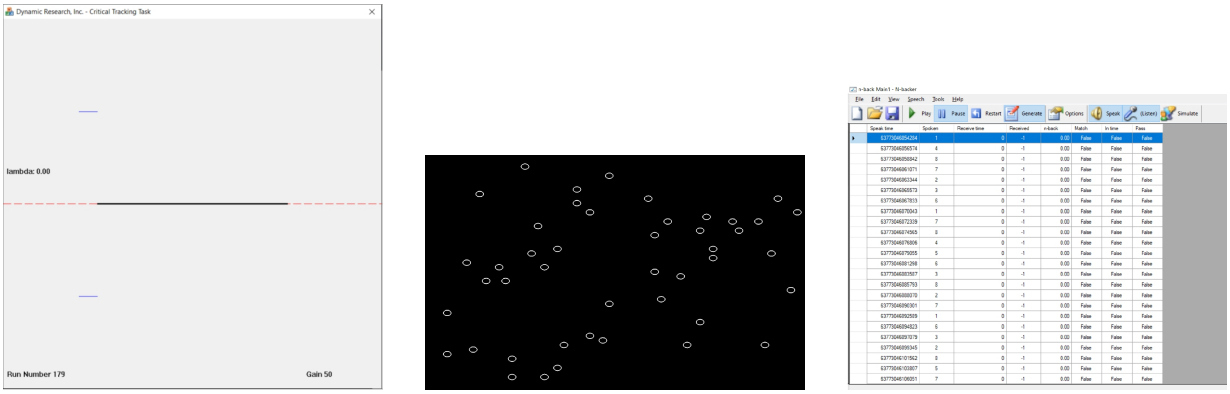


(a) Highway scenario.

(b) City road scenario.

Figure 3.1: Driving Scenarios.

Each participant experienced a different driving scenario in each trial, with the scenario order varied across participants. Each scenario began at a designated safe location. In the manual driving scenarios, the participants were asked to merge into the traffic and follow the verbal navigation from the experimenter. The experimenter guided participants along the route, and the scenarios ended when the vehicle reached a designated endpoint. In the automated driving scenarios, the participants were asked to enable the automation system at the beginning. During the autonomous control time, the semi-autonomous vehicle drove at 90 km/h on a two-lane highway and at 30 km/h on an urban, non-highway road, following the same trajectory as the verbal instruction. The automation system managed both



(a) CTT. (b) SuRT. (c) N-back Task.

Figure 3.2: Three types of NDRTs.

longitudinal and lateral control. Longitudinal control functioned similarly to an Adaptive Cruise Control system, adhering to speed limits, adjusting speed in response to slower leading vehicles, and maintaining the vehicle in the center of the current lane. All scenarios included moderate scripted traffic. On the highway, there were 10 vehicles per kilometer per lane, positioned roughly 100 meters from the ego vehicle. In the urban environment, traffic consisted of 6 vehicles per kilometer per lane, along with 2 pedestrians and 2 cyclists, also approximately 100 meters away from the ego vehicle.

In each scenario — both manual and automation — a hazard event was placed approximately 20 seconds before the designated endpoint. These hazard events were designed based on the driving environment, including situations such as vehicle collisions, pedestrian violations, or roads blocked caused by barriers, etc. The participants were expected to avoid the hazard event and proceed without stopping.

In the automation scenarios, the TOR would be issued 7 seconds before the ego vehicle would have collided with the hazard scene, and that would be about 2-3 minutes after the autonomous driving period started. Although no actual takeover occurred in the manual driving scenarios, the same "7-seconds-before-collision" spot was recorded as "nominal TOR point" for consistency in further analysis.

The TOR icon is shown in Figure 3.3. Participants were instructed to move immediately after they noticed the TOR. The participants' takeover actions were identified with their control move, detected by any movement of the steering wheel or pedals. Drivers were instructed to not touch the steering wheel or pedals during the autonomous control. If the participant did the takeover successfully and continued the driving task, the scenario



Figure 3.3: Takeover Request Icon

ended in the same designated spot as the manual scenario. If the participant failed to do a takeover, the vehicle gradually came to a stop after the TOR disappeared.

The experiment timelines are shown in Figure 4.8. The time intervals between key events during the trials are marked on the timelines. Figure 4.8a demonstrates an overall timeline for manual scenarios and Figure 4.8b shows an overall timeline for automation scenarios. This figure also illustrates the metrics used to measure SA and takeover performance, as well as the division of different TWs, which will be elaborated in the following Section 4.2.1 and Section 4.3.1.

3.1.2 Driving experiment instruments

The driver's eye movement was tracked by the Ergoneers Dikabilis 3 eye tracker (see Figure 3.4), which is a wearable wired binocular eye tracking glasses set with a 60 Hz tracking frequency and 0.05-degree visual angle accuracy. The eye camera resolution is up to 648 $\times$ 488 pixels, and the scene camera is 1920 $\times$ 1080 pixels. The eye tracker needs to be calibrated to suit the current wearer before each use.

The driving scenarios were programmed using CARLA (version: 0.9.12) [41], an open-source driving simulator software for autonomous driving. It provides open digital assets, such as urban layouts, buildings, and vehicles, that can be used freely. The eye tracker integrated software (D-Lab) and the CARLA were installed on a CyberPower PC Gaming Desktop. The simulator also included a GTR racing seat, a Logitech G29 steering wheel,



Figure 3.4: Ergoneers Dikabilis 3 eye tracker

and a set of pedals. The driving scenarios were displayed on one 27-inch full HD 1080p monitor, including rear and side mirrors. The NDRTs were presented on a Dell Precision 5520 laptop with a 15-in screen. The overall experiment setup is shown in Figure 3.5.

3.1.3 Driving experiment participants

3.1.3.1 Participants screening criteria

Participants must possess a valid Canadian driving license (*e.g.*, Ontario G2 or G) and have at least one year of driving experience to be eligible for the current driving experiment. Two screening tests were implemented before the formal driving part of the experiment to ensure that each participant had adequate cognitive ability. Test A was the Trail Making Test [27], designed to assess participants' visual attention and task switching. The test-taker needed to connect a sequence of 25 consecutive targets on the computer screen, which could be numbers from 1 to 25 or numbers from 1 to 13 and letters from A to L, and the cut-off complete time for these two tests is 29 seconds and 273 seconds or greater respectively. Task B was the Sickness Susceptibility Questionnaire-Short [57] (Appendix C), which was used here to exclude the participants who would have higher risks of motion sickness, and the cut-off score for this test was 23. The participants must also have a visual acuity better than 20/50 using a standard Snellen eye chart with both eyes open, with or without the



Figure 3.5: Experiment environment setup

aid of corrective lenses.

Most participants were recruited from the local Waterloo and Kitchener community. The young drivers were primarily students from the University of Waterloo and other local universities. The middle-aged drivers were approached via posters and social media advertisements. The older drivers were contacted with the assistance of the WRAP (Waterloo Research Aging Participant Pool), who provided the contact information of older people who have registered and explicitly noted interest for future university experiments.

3.1.3.2 Participants demographics and sample size

After screening, a total of 82 participants were recruited; however, 10 data points were excluded from the final analysis due to simulator sickness or unexpected equipment failures. In the end, we completed the complete experiment with a balanced sample of 24 young drivers (18-24 years old), 24 middle-aged drivers (35-55 years old), and 24 older drivers (65 and plus years old) and collected all required data. Young drivers had a mean age of 21.15 years old (SD = 3.35), middle-aged drivers had a mean age of 42.04 years old (SD = 3.35), and older drivers had a mean age of 80.4 years old (SD = 6.14). In total, there are 32 females (11 young, 11 middle-aged, and 10 older) and 40 males (13 young, 13 middle-aged, and 14 older).

Although driving experience is not a core study objective in the current thesis, we report

Table 3.1: Participants' age groups

Age Group	Number	Mean Age (SD)	Female	Male
Young (18–24)	24	21.15 (3.35)	11	13
Middle-aged (35–55)	24	42.04 (3.35)	11	13
Older (65+)	24	80.40 (6.14)	10	14
Total	72	–	32	40

participants' driving experience here for reference. Younger drivers had an average driving experience of 4.3 years (SD = 1.8) and an average driving mileage of 35,307.6 kilometers (SD = 56,420.9). Middle-aged drivers had an average driving experience of 22.3 years (SD = 5.9) and an average driving mileage of 712,423.1 kilometers (SD = 2,046,081.1). The large standard deviation for mileage in this group reflects large variation and uncertainty in self-reported mileage amounts. Older drivers had an average driving experience of 59.7 years (SD = 10.1). During the experiment, we found that it was very difficult for older drivers to accurately report their lifetime driving mileage, so this information was not collected for this group.

3.1.4 Driving experiment procedures

The consent of all participants was obtained before the commencement of the experiment. Firstly, they took the Trail Making Tasks, the MSSQ-Short survey, and the visual acuity test to make sure they were qualified for the experiment. Then, they took a short demographics questionnaire (Appendix B) to indicate their age, driving experience, and knowledge about autonomous driving. No private information was collected here. Following the questionnaire, they were given verbal instructions about the study objectives and their tasks, which included the automated driving system, the drivers' responsibilities in the takeover process, and how to engage in the NDRTs. They were notified that the autonomous system can handle the driving tasks until it gives the TOR, so they do not need to keep monitoring the screen and should try to perform well in the NDRTs. They confirmed that upon receiving a TOR, they would immediately take control of the vehicle and take any necessary actions, based on their own judgment, to ensure driving safety. They were also instructed to follow traffic rules at all times throughout the experiment.

The participants were then given a practice session to ensure that they were familiar with the steering wheel and pedals of the simulator before the formal experiment session. The practice session used a separate driving scenario from the formal experiment to avoid learning effects and ensured that the participants were unfamiliar with the experimental

scenarios. No takeover events were included in the practice session, so participants would not develop rapid takeover responses in advance. However, the takeover procedure was explained to them verbally to ensure they understood what to expect during the experiment.

Participants wore the eye tracking device and completed a calibration process prior to the start of the formal experiment session. The formal driving session began after the participants felt confident enough to handle the simulator. They performed 12 driving scenarios, each lasting 5 minutes on average. The total duration of the experiment is about 90 minutes on average. Finally, the participants were remunerated for their time and thanked for their participation.

3.2 The piloting experiment

The experimental protocol was approved by the Office of Research Ethics of the University of Waterloo (ORE # 46113).

3.2.1 Piloting experiment design

A series of simulated landing scenarios, representing the descending process from the air to the runway, were designed in the piloting experiment. The takeover tasks in the current experiment were necessitated by a series of malfunction conditions, requiring pilots to first detect the malfunction and then assume manual control of the airplane. The landing scenarios took place at the Region of Waterloo International Airport (CYKF) runway 26, starting at 3,000 ft. At that point, the airplane was already on approach and had aligned with the runway centerline (LOC intercepted). A side view of the landing trajectory is shown in Figure 3.6. From the y-axis, it can be observed that the airplane started at an altitude of 3,000 ft and descended to land on runway 26, which has an approximate elevation of 1,000 ft above sea level (ASL). All scenarios started with autopilot engaged, during which no manipulation was required except for managing the engine power. The pilots were required to complete several landing processes and handle any possible malfunction situations during the process. A 90-degree crosswind at 10 knots was set throughout all scenarios. This could cause the airplane to roll left from the runway centerline, potentially leading to lateral misalignment if not properly corrected during the approach.

To enhance the realism of the scenarios and generalizability of the conclusions, distinct landing situations were included in the scenarios. Similar to the driving experiment, the variations were not designed as formal experimental manipulations but served to enhance

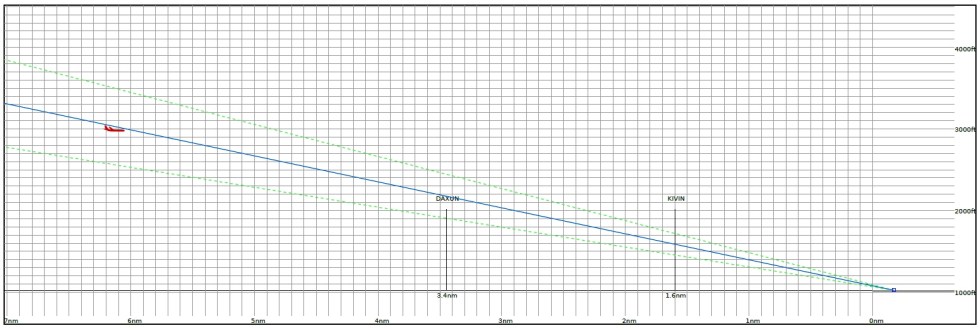


Figure 3.6: Landing trajectory

the ecological validity of the simulation. To provide diverse piloting conditions, three types of malfunction were included: no failure, LOC failure, and autopilot failure. In the no failure scenarios, the pilots landed the aircraft without any interruption or malfunction. The pilots were informed that the autopilot must be disengaged at a minimum altitude of 250 ft above ground level (AGL), requiring them to switch to manual control at that point, regardless of whether a failure had occurred. This rule was implemented to simulate standard operating procedures, ensure sufficient time for manual landing maneuvers, and avoid overreliance on automation during the critical final phase of the approach. The LOC failure disabled the radio beam sensor of the lateral direction, so the airplane would land following a wrong lateral direction. The autopilot failure disabled the whole automated flight control system, so the airplane would lose both lateral and vertical direction control. Since the existence of crosswind, the aircraft would face misalignment in both failure types if not controlled properly. Two weather conditions were included in the simulation: clear weather and low visibility weather. In the low visibility conditions, the airplane was still safe to land based on the instrument approach procedure charts of runway 26 of CYKF. Figure 3.7 shows these two weather conditions in the front windows when the airplane was near the decision height. Half the scenarios included ATC communication, so the pilots needed to report their landing normally as an approach in real life. The scripts used for ATC are attached in the Appendix D. In the no ATC scenarios, pilots did not need to report anything.

The combinations of scenarios above were randomized, and participants experienced a different scenario in each trial. In the scenarios with failures, the failures were manually triggered near DAXUN — the final approach fix (FAF) waypoint associated with CYKF, as shown in Figure 3.6). This occurred approximately two minutes after the scenarios started depending on the speed controlled by the pilots. Failure alerts appeared in the center of the instrument panel, with distinct indications between two types of failures. Figure 3.8



(a) Clear weather.



(b) Low invisibility.

Figure 3.7: Two weather conditions in piloting experiment.

demonstrated the instrument panel display under three conditions: normal operation with autopilot engaged (Figure 3.8a), a LOC failure indicated by the disappearance of the diamond symbol on the right side of the heading gauge (Figure 3.8b), and an autopilot failure alert shown at the top of the attitude gauge (Figure 3.8c). Pilots were expected to notice the alerts, disengage the autopilot (for both failure types) and continue to land manually. Their takeover actions were identified at the moment they manually disengaged the autopilot by double-clicking the red button located to the left of the control yoke. In the no failure scenarios, pilots also needed to switch to manual landing at 250 ft AGL. Therefore, all scenarios were ended with manual landing.

Pilots from two experience groups (with or without the Instrument Rating) were recruited to participate in the experiment. More details will be covered in Section 3.2.3.

The experiment timelines are shown in Figure 4.9. The time intervals between key events during the trials are marked on the timelines. Figure 4.8a demonstrates an overall timeline for no failure scenarios, and Figure 4.8b demonstrates an overall timeline for scenarios with failures. This figure also illustrates the metrics of TOT, as well as the division of different TWs, which will be elaborated in the following Section 4.2.2 and Section 4.3.2.

3.2.2 Piloting experiment instruments

This study was conducted on the ALSIM AL250 flight simulator administered by the Waterloo Institute for Sustainable Aeronautics (WISA), and the setup of the experimental



(a) No failure.



(b) LOC failure.



(c) Autopilot failure.

Figure 3.8: Instrument panel with three types of failure conditions.



Figure 3.9: Piloting experimental environment setup

environment is shown in Figure 3.9. The simulator consists of two primary components: a physical flight cockpit and a computer system running flight simulation software. The cockpit setup included realistic flight controls, such as a control yoke, rudder pedals, and throttle, providing tactile feedback and immersive interaction. The simulation software was configured to replicate the flight dynamics and cockpit layout of a single-engine aircraft — Cessna 172. To reflect traditional training environments, the simulated cockpit featured an analog instrument panel, comprising the standard "six-gauge" flight instruments (airspeed, attitude, altimeter, roll, heading, and vertical speed), rather than a glass cockpit display.

The participants' eye movements were tracked by the AdHawk MindLink eye tracker from AdHawk Microsystems Inc., Waterloo, ON, Canada. This is a pair of wearable eye tracking glasses (see Figure 3.10) with a 60 Hz tracking frequency and 1920 x 1080 pixels scene camera resolution. The AdHawk software can provide detection of eye movement events (e.g. fixation and saccades) and AOI locations. This was achieved by detecting the locations of the QR code markers placed on the simulator dashboard, which are covered

with special glass layers, aiming at minimizing the interference with pilots' operation. This eye tracker needs to be calibrated to suit the current wearer before each time of usage.



Figure 3.10: AdHawk eye tracking glasses

3.2.3 Piloting experiment participants

3.2.3.1 Participants screening criteria

Two groups of pilots were recruited, one group with their Instrument Rating (Expert group) and the other group without (Learner group). Each group has 11 participants. The Instrument Rating authorizes pilots to conduct approaches using the ILS independently. All participants were required to hold a valid Private Pilot License (PPL).

3.2.3.2 Participants demographics and sample size

The Expert group (Instrument Rating holders) had an average of 964.69 flight hours (SD = 1028.95), while the Learner group (non-Instrument Rating holders) had an average of 153.74 flight hours (SD = 30.61). The Learner group consisted of 3 females and 8 males, whereas the Expert group included 1 female and 10 males.

3.2.4 Piloting experiment procedures

The consent of all participants was obtained before the commencement of the experiment. Firstly, they took a short demographics questionnaire to indicate their age, flight hours,

Table 3.2: Participants' experience group

Experience group	Number	Flying hours (SD)	Female	Male
Learner	11	153.74 (30.61)	3	8
Expert	11	964.69 (1028.95)	1	10
Total	22	–	4	18

processed certificates, and their knowledge about ILS. The questionnaire is attached in Appendix F. No private information was collected here. Following the questionnaire, they were given verbal briefing about the purpose of the study and their tasks regarding the current experiment. A briefing board was used here to effectively deliver relevant information (see Figure 3.11). Pilots did a few rounds of ILS landing in clear weather conditions with no system failure included and no extra communication with the ATC as practice. After the pilots felt confident to handle the simulator, the formal experiment started. They were asked to wear the eye tracker throughout the whole experiment session, and they needed to conduct the calibration again between trials if necessary. They performed 12 landing scenarios and each one took about 5 minutes on average. The total duration of the whole process was about 120 minutes. Finally, the participants were remunerated accordingly and thanked for their participation.

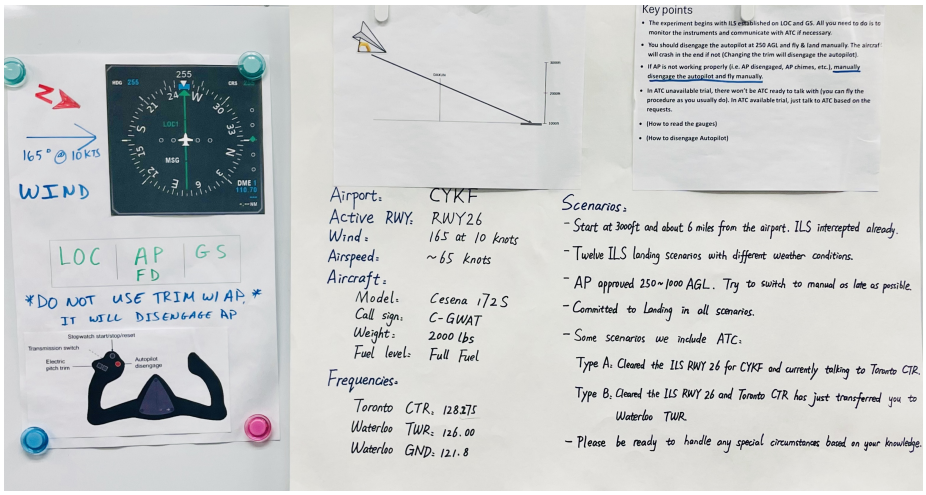


Figure 3.11: Briefing board before the formal experiment

Chapter 4

Methodology

This chapter introduces the data analysis methodologies used for both driving and piloting domains, including eye tracking data pre-processing, metrics used to measure SA and takeover performance, TW separation criteria, and an overview of data processing methods. Figure 4.1 outlines the three topics of this thesis and illustrates the connections between eye-tracking data and various variables. This structure was applied to data collected from both the driving and piloting experiments, with minor differences in the metrics explored. The current study implemented all data pre-processing, t-tests, and machine learning models using Python (version 3.9.21).

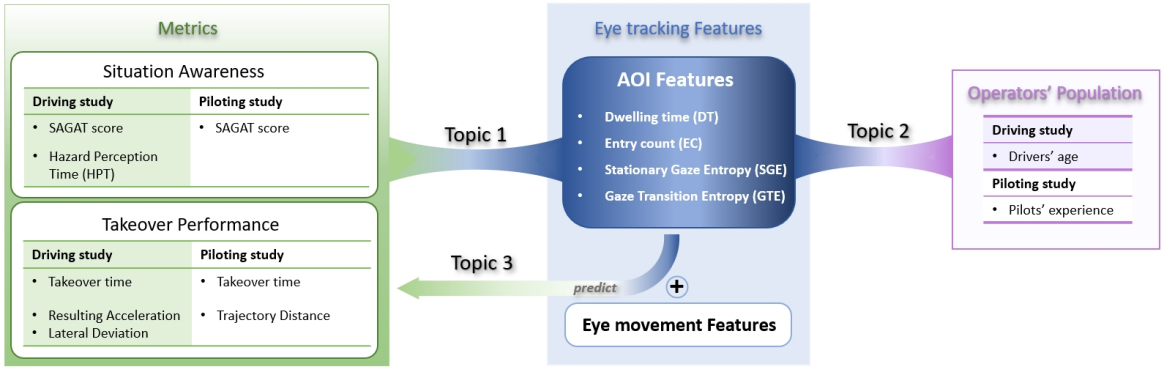


Figure 4.1: Data analysis structure

As shown in the Figure 4.1, eye tracking features are the core of the current thesis and all three topics are related to them. The eye tracking features include two categories of

features, AOI features, which were associated with where the gaze was located, and eye movement features, which linked predominantly to fixation and saccades. The SA-related metrics and performance-related metrics were calculated for both driving and piloting domains. Those metrics are shared by both Topic 1 and Topic 3. In Topic 2, the operators' population is a pre-existing grouping, which was collected through the demographic questionnaires prior to the start of the experiment.

The three topics in the current thesis are:

Topic 1: Differences in eye movement between operators with different Situation Awareness levels and takeover performance levels in a conditionally automated system.

Topic 2: Differences in eye movement between operators from different population groups during the takeover process in a conditionally automated system.

Topic 3: Using eye tracking features to predict operator's Situation Awareness and takeover performance in a conditionally automated system with machine learning methods.

Topic 1 and Topic 2 focused on the AOI feature, which reflects spatial gaze distributions. These features are intuitive to interpret and hold potential for future training or instructional applications. In addition to the AOI feature, Topic 3 incorporates additional eye movement features to comprehensively capture gaze behavior for input into machine learning models. Section describe the metrics grouping criteria and relevant references used in the current thesis. It also introduced the

Section 4.1 introduced the pre-processing steps for eye tracking data, including the definition of eye movement feature and AOI feature. Furthermore, the steps of extracting the AOI feature are discussed separately across the driving and piloting domains, due to differences in experimental setups and eye tracking equipment. Section 4.2 introduced the metrics used to evaluate operators' SA and takeover performance in two domains. Section 4.3 introduced the TW segmentation methods in two domains. Section 4.4 introduced the statistical approach, including the MANOVA and Welch's t-tests with FDR, used in Topic 1 and Topic 2, as well as the the machine learning method pipeline used in Topic 3.

Together, these sections form the methodological foundation of this thesis. The corresponding results for both the driving and piloting studies—organized by Topic 1 through Topic 3 — are presented in Chapter 5 to 10.

4.1 Eye tracking data pre-processing

In both driving and piloting studies, eye tracking data was collected in each trial of the experiment. Participants wore the eye tracking glasses throughout the entire experiment. To gain insights into their eye movements in conditionally automated systems, a series of eye tracking features were extracted and used across all three topics discussed in this thesis. Based on the review in Chapter 2 (Table 2.2), the current thesis chose to use two categories of eye tracking features: the eye movement feature and the AOI feature.

4.1.1 Eye movement feature

The first category of eye tracking features is the eye movement feature. Table 4.1 below lists all 12 eye movement features used in the current thesis. The eye movement features selected here are the combinations of the total number, average value, and standard deviation of operators' fixations and saccades. Fixations and saccades, as two fundamental types of eye movements, have shown a high correlation with operators' SA and performance in prior studies. To capture the participants' visual behavior as thoroughly as possible, we included a comprehensive set of eye movement features.

The process of classifying eye movements into fixations and saccades — commonly referred to as event detection — was completed using the native software associated with the eye trackers used in each experiment. In the driving experiment, the Ergonees Dikablis 3 eye tracking glasses were integrated with D-lab software. This software allows users to configure the minimum fixation duration for event detection, and the default threshold of 120 milliseconds was applied here. In the piloting experiment, the eye movement was captured by the AdHawk eye tracking glasses with the AdHawk Hub software. The company also provided its own event detection as well as AOI detection algorithms. Across both systems, eye movements were identified as either saccades or fixations for the majority of time. Other eye movement types, such as blinks and smooth pursuits, were excluded from the analysis due to their minimal presence in the dataset and limited relevance to the current research objectives. The eye tracking feature used in both driving and piloting experiments shared same definition and calculation procedures, so no further domain-specific discussions are necessary here.

Table 4.1: Eye movement features

Num	**Eye movement feature**	**Explanation**
1	num_sac	Number of saccades
2	ave_sac_len	Average saccade length
3	std_sac_len	Standard deviation of saccade length
4	ave_sac_ang	Average saccade angle
5	std_sac_ang	Standard deviation of saccade angle
6	max_sac_ang	Maximum saccade angle
7	num_fix	Number of fixations
8	ave_fix_len	Average fixation length
9	std_fix_len	Standard deviation of fixation length
10	max_fix_len	Maximum fixation length
11	ave_pupil	Average pupil diameter
12	std_pupil	Standard deviation of pupil diameter

4.1.2 Area-of-Interest Features

The second category of eye tracking features is the AOI feature, which describes eye movements in terms of gaze location within pre-defined AOIs. Table 4.2 lists all four AOI features used in the current thesis, including Dwelling time, Entry count, SGE and GTE.

The Dwelling time reflects how much visual attention a participant allocates to a specific area. A longer Dwelling time generally indicates more attention allocated. For instance, if a pilot spends more time in the instrument panel, it may suggest focused monitoring or uncertainty requiring verification. Dwelling time is usually defined as the total time participants spend fixating at each pre-defined AOI. Since the sampling frequency of the eye trackers was unstable, it is hard to get the exact time spent in each AOI. Considering only the relative time length spent on different AOIs matters, the Dwelling time for a given AOI is defined as the percentage of time the participant spent in that AOI relative to the total gaze duration within the analyzed period in the current thesis. In the current thesis, Dwelling time features—part of the AOI feature set—are named after their corresponding predefined AOIs, followed by "_DT" to indicate "Dwelling time" (e.g., *Mirror_DT* or *Attitude_DT*).

Entry count refers to the total number of time that the participant's gaze moves from outside to inside of an AOI. Unlike Dwelling time, which measures how long attention is maintained, entry count reflects how frequently an AOI is revisited. A higher entry count often reflects increased monitoring or scanning behavior, suggesting that the participant

Table 4.2: AOI feature

AOI feature	**Explanation**	**Calculation** (From Gaze-to-AOI Table and Gaze Transition Table)
Dwelling time (for each AOI)	The percentage of time the participant spends in the AOI among the total time length.	Number of ones divided by the total number of rows
Entry count (for each AOI)	The total number of time that the participant's gaze moves from outside to inside of an AOI.	Number of contiguous groups of ones
Stationary gaze entropy (SGE)	It is a probability distribution of gaze locations, quantifying the average uncertainty level of a sequence of fixations.	$H(x) = -\sum_{i=1}^{n}(p_i)log_2(p_i)$ (1)
Gaze transition entropy (GTE)	It represents the probability distribution of gaze transitions between AOIs, quantifying the average uncertainty in the sequence of gaze shift.	$H(x) = -\sum_{i=1}^{n} p_i \sum_{j=1}^{n} p(i,j)log_2 p(i,j)$ (2)

repeatedly checks that area. In contrast, a lower entry count may indicate either confidence in the information from a single glance or a lack of perceived relevance. In the current thesis, the Entry count is defined as the number of fixation entries into the given AOI. In the current thesis, Entry count features—part of the AOI feature set—are named after their corresponding predefined AOIs, followed by "_EC" to indicate "Entry count" (e.g., *Mirror_EC* or *Attitude_EC*).

The concept of entropy has gained more and more attention in examining gaze behavior quantitatively in recent years. This concept attempts to describe the average information or uncertainty associated with choices [141]. There are two types of entropy used in the current thesis, the SGE [141] and the GTE[25]. The SGE depicts the general uncertainty of gaze allocation in an experimental trial which can be calculated by the equation (1) in Table 4.2. The p_i in the equation refers to the possibility of a gaze landing in the AOI i. The GTE depicts the uncertainty of gaze transfer between AOIs, which can be calculated by the equation (2) in Table 4.2. The $p(i, j)$ in the equation refers to the possibility of a gaze transfer from AOI i to AOI j. In both equations, the n is the number of predefined AOIs. In the current thesis, the term "gaze" used in the definitions of SGE and GTE above specifically refers to fixations, meaning only fixations are considered valid gaze points. A higher entropy reflects greater uncertainty in eye movements, typically associated with behaviors such as searching or monitoring. In contrast, lower entropy indicates more focused and stable gaze patterns, often concentrated on a single location.

4.1.3 AOI feature calculation

In preparation for calculating the final AOI features, a key intermediate step is gaze-to-AOI mapping. This step identifies which AOI the gaze point falls into at each fixation and produces the Gaze-to-AOI Table as well as the Gaze Transition Table. The Gaze-to-AOI Table is a one-hot encoded matrix, where each row corresponds to one fixation and each column represents a pre-defined AOI. A value of 1 in each row indicates that the fixation falls into that corresponding AOI at that timestamp. An example of a Gaze-to-AOI Table is shown in Table 4.3.

Table 4.3: Gaze-to-AOI Table using One-Hot Encoding

rec_time	Gaze_X	Gaze_Y	AOI 1	AOI 2	AOI 3	AOI 4	.....	AOI N
timestamp 1	1003.667	630.253	0	0	0	1	.....	0
timestamp 2	1004.558	636.069	0	1	0	0	.....	0
timestamp 3	1003.978	628.017	0	1	0	0	.....	0
timestamp 4	1004.205	636.782	0	1	0	0	.....	0
timestamp 5	1002.725	627.832	0	1	0	0	.....	0
timestamp 6	1004.117	635.661	0	0	0	1	.....	0
timestamp 7	1002.104	626.048	0	0	0	1	.....	0
timestamp 8	1000.688	636.149	0	0	0	1	.....	0
timestamp 9	1002.784	626.95	0	0	0	1	.....	0
timestamp 10	1004.055	633.328	0	0	0	1	.....	0
.....	.....	.....	.....	.....	.....	.....	.....	.....

The Gaze Transition Table is an N by N matrix, where N is the number of AOIs in a specific simulation environment. It is derived from the Gaze-to-AOI table, by recording each transition between pairs of AOIs. Each cell in the matrix indicates the total number of transitions from one specific AOI to another. Table 4.4 is an example of the Gaze Transition Table.

Table 4.4: Gaze Transition Table

	AOI 1	AOI 2	.....	AOI N
AOI 1	/	a	.....	b
AOI 2	c	/	.....	d
.....	.....	.....	/	.....
AOI N	e	f	.....	g

The "Calculation" column of Table 4.2 in the Section above showed how to get the values of Dwelling time, Entry count, SGE and GTE from the Gaze-to-AOI Table and Gaze Transition Table. The dwelling time of a certain AOI is calculated as the number of ones divided by the total number of rows in the corresponding column of that AOI in the Gaze-to-AOI Table. The Entry count of a certain AOI is calculated as the number of contiguous groups of ones in the corresponding column of the given AOI, representing the number of times the gaze entered the AOI from outside. The SGE is calculated using equation (1), where p_i is the possibility of a gaze landing in AOI i, which equals the dwelling time value of the AOI i. The GTE is calculated using equation (2), where the $p(i, j)$ is the possibility of a gaze transfer from AOI i to AOI j. This is determined by dividing the number of transitions from AOI i to AOI j (i.e., the value in row i, column j of the Gaze Transition Table) by the total number of transitions from AOI i to all other AOIs (i.e., the sum of all values in row i of the Gaze Transition Table).

However, the positions of the AOIs were dynamic in the eye tracking videos. In the current driving and piloting experiments, eye tracking glasses were mounted on the participant's head, and the front scene camera captured the whole front view of the wearer, which kept moving dynamically with the participant's head throughout the experiment. Since that pre-defined AOIs in both driving and piloting experiments were fixed areas inside the simulation environment, they also moved dynamically in the eye tracking video captured by the front camera. To determine which AOI each gaze point fell into, a transfer process was required to align the gaze coordinates with the dynamic position of the AOIs.

The transfer processes were achieved by the QR code markers, which were introduced in Chapter 3.1.2 and 3.2.2 before. The QR code markers were placed in the simulator to mark the locations of AOIs in the participant's front view because their coordinates can be automatically identified by the eye tracking software. Given the spatial relationships between the QR code markers and the pre-defined AOIs were fixed, the coordinates of the markers and the participant's gaze coordinates can determine where they were actually fixated at.

The procedures described above outline the overall approach for Gaze-to-AOI mapping. However, due to differences in AOI definitions, eye tracking glasses, and simulator setups between the driving and piloting domains, the preceding steps for deriving the Gaze-to-AOI Table and Gaze Transition Table will be discussed separately in the following sections.

4.1.3.1 AOI feature calculation in driving

In the driving experiment, the six pre-defined AOIs were selected based on previous studies as well as current research objectives. These AOIs include the Rear-view mirror, Dashboard, Right mirror, Left mirror, Center, and Others (see Figure 4.2). The first four AOIs are common areas that contain critical information in a normal vehicle, such as mirrors and the dashboard. The AOI "Center" area refers to the middle portion of the road ahead, where the TOR icon appears. This area approximately matches the size of the TOR icon and typically aligns with the center of the driver's forward view of the road, which is also where accidents and hazards most commonly occur. The AOI "Other" indicates the whole view captured by the front camera of the eye tracking glasses except the five AOIs above.

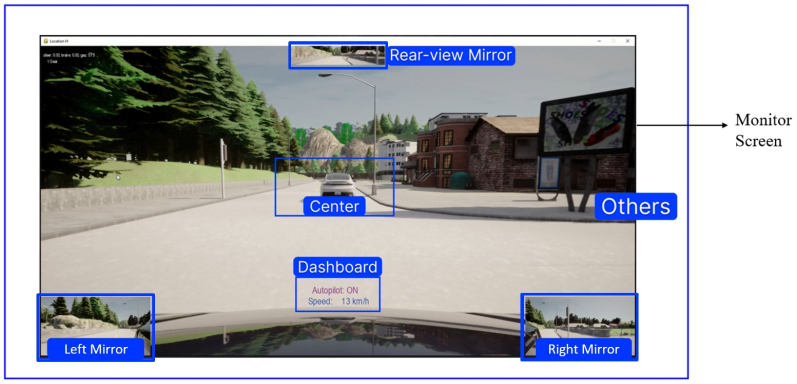


Figure 4.2: Driving experiment AOIs

As shown in Figure 4.2, the software window of a driving scenario was maximized and displayed on a 27-inch monitor screen. In the experiment, four QR code markers were placed next to the four corners of the monitor (see Figure 4.3), which can be automatically recognized and tracked by the D-Lab software with the same frequency of other eye tracking variables.

At each timestamp, the gaze point's coordinates (shown as the red dot in Figure 4.3), as well as the markers' coordinates, were recorded. The four QR code markers, forming a quadrilateral shape in Figure 4.3, correspond with the four corners of the scenario window in Figure 4.2, which forms a rectangle. Figure 4.4 demonstrates the transfer process [38]. In this Figure, image 1 (a rectangle) and image 2 (a quadrilateral) share an affine transfer relationship. By selecting three corresponding points from each image, a transfer matrix M can be produced. Once the M is obtained, any points on image 2 can be mapped back to its corresponding location on image 1 by left-multiplying it with the inverse of

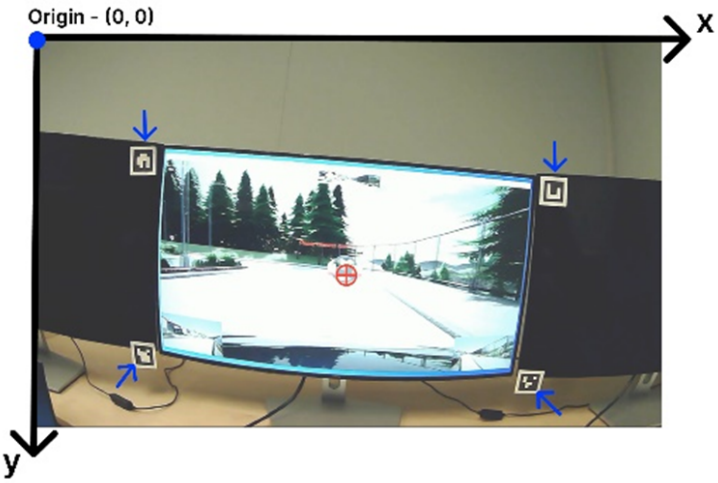


Figure 4.3: Driving experiment markers in the scene camera frame

M. For the driving experiment, the corners of the scenario window correspond to image 1 and the markers' coordinates correspond to image 2. Once the M is obtained, the eye gaze coordinates can be multiplied by the inverse of M to determine the corresponding gaze location on the monitor screen in Figure 4.2. Determining the corresponding gaze location on the monitor screen provides sufficient information to identify which AOI the gaze location falls into, as the relative locations between different AOIs and the monitor screen are fixed. Then, the Gaze-to-AOI Table was obtained and the following steps of calculating AOI feature were conducted. Also, four — rather than three — QR code markers were placed to increase the reliability of the method because they may not be detected by the software occasionally.

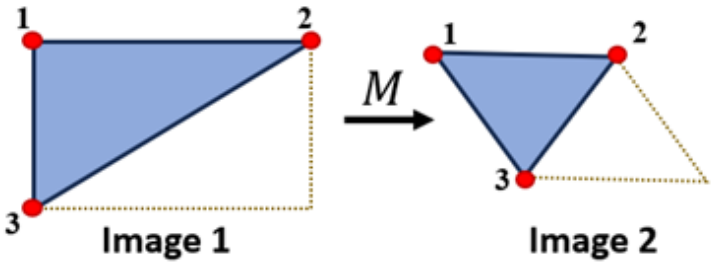


Figure 4.4: Affine Transformation in the driving experiment

4.1.3.2 AOI feature calculation in piloting

In the piloting experiment, nine AOIs are defined: Front, representing the front window of the airplane; Power, referring to the aircraft's RPM gauge; six AOIs inside the Instrument Panel, including Airspeed, Attitude, Altitude, Roll, Heading, and Vertical speed; and Others, which includes all areas captured by the front camera of the eye tracking glasses except the eight AOIs above (see Figure 4.5).

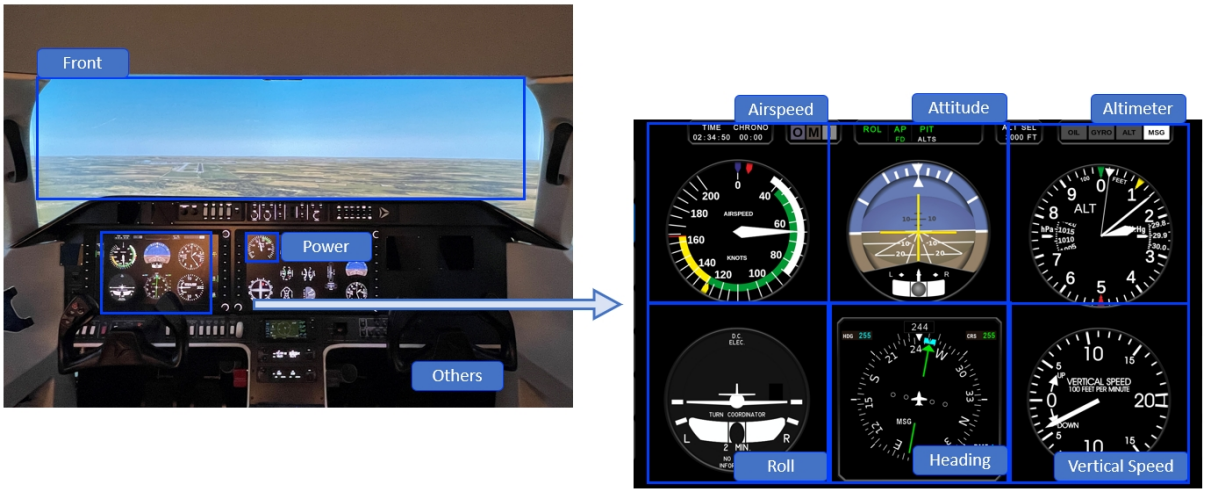


Figure 4.5: Piloting experiment AOIs

Five QR code makers were placed on the cockpit, but only three of them were used eventually, as marked in Figure 4.6. The QR code markers were covered by a piece of special glass to make them look dark from the outside with the naked eye, minimizing interference with pilots' operations [11]. Figure 4.7 illustrates the visual differences of the markers with and without the glass layers. Meanwhile, the AdHawk eye tracking device can still automatically recognize and track the QR code markers' locations. At each timestamp of the eye tracking file, the gaze point coordinates, as well as the markers' coordinates, are recorded. Most of the time during the piloting experiment, only the left two markers were captured by the front scene camera of the eye tracking device because only one pilot sat pretty close to the instrument panel and barely looked to the side during the experiment.



Figure 4.6: Piloting experiment markers in the scene camera frame

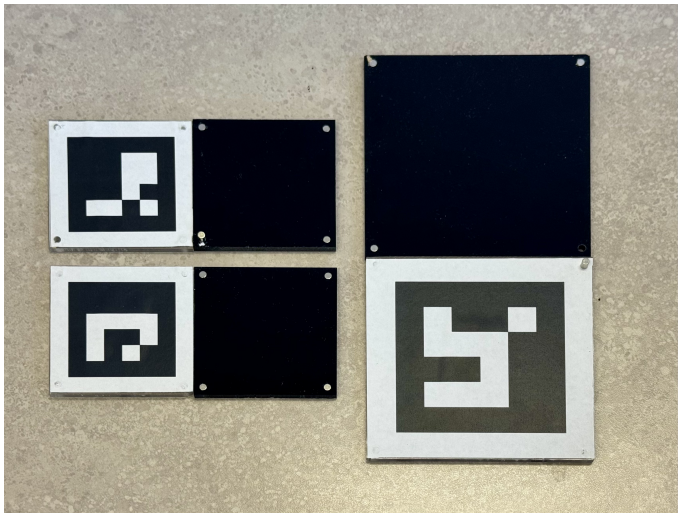


Figure 4.7: Comparison of marker visibility with and without glass layers

A 3D-transferring algorithm was provided by the AdHwadk eye tracking company, so the Gaze-to-AOI Table were obtained directly from the piloting experiment raw data with minor processing, and the AOI features were calculated from there.

4.2 Situation Awareness and Takeover performance Metrics

Operators' SA and takeover performance are two key constructs examined in this thesis. They are measured by different metrics due to different setups of the experiment scenarios. The following sections present these measurements separately for the two domains.

4.2.1 Driving Study

Table 4.5 summarized all metrics used to measure drivers' SA and takeover performance. The following sections will introduce the details of these measurements.

Table 4.5: Evaluated Metrics for SA and Takeover Performance in the driving experiment

Key Constructs	**Evaluated Metrics**
Situation Awareness	SAGAT score
	Hazard Perception Time
Takeover Performance	Takeover Time
	Resulting Acceleration

4.2.1.1 Drivers' Situation Awareness

Drivers' SA are measured from two aspects. One is the score they got from the direct SA measurement: SAGAT. As a freeze probe technique, SAGAT is one of the most popular direct SA measurements in previous driving research. It is also suitable for the current driving study because the drivers are required to pay attention to surrounding information throughout the experiment, which creates a chance for researchers to design various SAGAT questions. All questions used in the experiment are listed in the Appendix A. After careful consideration of the advantages and disadvantages, a delay-SAGAT method was employed to measure the drivers' SA. This method employed the SAGAT questions at the end of each trial to avoid the intrusion of the experiment process, which has been used in a previous driving study [131]. Endsley's research [110] validated the delayed-SAGAT by proving pilots' consistent accuracy in responding to SAGAT queries across a 6-min period after a freeze, indicating no memory decay. Therefore, in the current driving experiment, two SAGAT questions were asked at the end of each driving scenario. Drivers' answers to

the two questions sum up their score for that trial. Each SAGAT question was scored as correct (1 score), incorrect (0 score), or partially correct (0.5 score), so the driver could get up to 2 scores for each trial.

The other SA measurement was the HPT. HPT has been showed to be an effective indicator of SA in previous research [22]. As introduced above, a hazard scene was placed at each scenario, and it is critical for drivers to visually perceive the hazard location to ensure safety. Therefore, HPT was considered another metric to represent SA. Since drivers had not yet taken any evasive actions when they first fixated on the hazard scene, their HPT reflects their perceptual awareness rather than their behavioral response, and thus falls under the scope of SA rather than takeover performance. The HPT was defined as the interval between the TOR (or "nominal TOR point" in manual scenarios) to the moment when the hazard scene was fixated on, which was identified from drivers' eye tracking videos manually. The HPT is illustrated in the timelines for both types of scenarios in Figure 4.8.

4.2.1.2 Drivers' takeover performance

Driver's takeover performance was measured from two parts: TOT and Resulting acceleration. The TOT is the interval between the TOR and the moment when the driver conducts a control move, detected by any movement of the steering wheel or pedals. The TOT is illustrated in the timeline for automation scenarios in Figure 4.8b. The Resulting acceleration refers to the force transferred from the vehicle's tires to the ground, defined by the equation below:

$$ResultingAcceleration = \sqrt{MaxLongAcc^2 + MaxLatAcc^2}$$

In the equation, the maximum longitudinal acceleration ($MaxLongAcc$) is the maximum force applied during acceleration and braking, and maximum lateral acceleration ($MaxLatAcc$) is the maximum force that acts transversely to a vehicle's travel direction, which occurs during cornering [91]. A higher value indicates a more unstable takeover.

4.2.2 Piloting Study

Table 4.6 summarized all metrics used to measure pilots' SA and takeover performance. The following sections will introduce the details of these measurements.

Table 4.6: Evaluated Metrics for SA and Takeover Performance in the piloting experiment

Key Constructs	Evaluated Metrics
Situation Awareness	SAGAT score
Takeover Performance	Takeover Time
	Trajectory distance

4.2.2.1 Pilots' Situation Awareness

Pilots' SA were measured by the score they obtained on the direct SA measurement: SAGAT, which is suitable for the current piloting experiment since there was a substantial amount of information in the surrounding areas that required the pilots' attention. All questions used in the current piloting experiment are listed in the Appendix E. Similar to the driving experiment, a delayed-SAGAT method was implemented here as well, because it was very challenging to freeze the landing tasks and ask SAGAT questions during each trial. Two questions were asked at the end of each trial, and pilots' answers to the two questions sum up their score for that trial. Each SAGAT question was scored as correct (1 score), incorrect (0 score), or partially correct (0.5 score), so the pilot could get up to 2 scores for each trial.

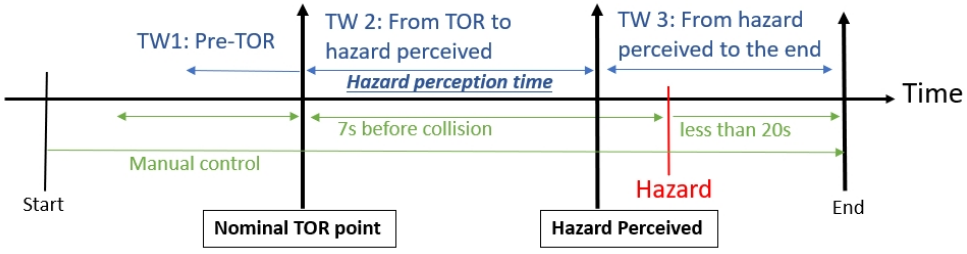
4.2.2.2 Pilots' takeover performance

Pilots' takeover performance was measured by two parts: TOT. The TOT is the interval between the alert and the moment when the pilot turns off the autopilot function. The TOT period is illustrated in the timeline for the scenarios with failures in Figure 4.9b.

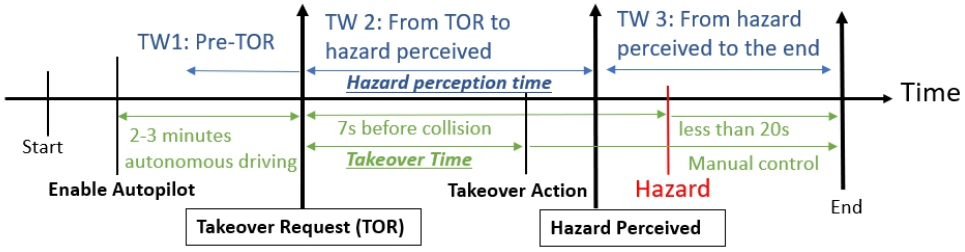
4.3 Time Window

The current thesis analyzed operators' eye movement across three different TWs, providing a comprehensive view of how their eye movement patterns change over time during the takeover process. Rather than arbitrarily extracting eye movement features from a few seconds before or after a TOR or malfunction alert, this work segmented each trial based on critical events, such as a TOR or takeover action. The events selected for the driving and piloting domains differ slightly due to the distinct nature of these two operations, and they will be discussed respectively below.

4.3.1 Driving study



(a) Driving experiment timeline for manual scenarios.



(b) Driving experiment timeline for automation scenarios.

Figure 4.8: Driving experiment timelines for two types of scenarios.

In the driving experiment, the total duration of each scenario varied slightly depending on drivers' operation speed and reaction time. Regardless of whether a scenario began with autonomous control or manual driving, it took about three minutes for the ego vehicle encountered the hazard scene in the front. Given that the entire scenario was relatively long and only certain periods were relevant to drivers' SA and takeover performance, specific TWs were extracted. All further analysis of drivers' eye movements are solely within those TWs. On the one hand, how drivers with better performance behaved during different TWs can be highlighted. On the other hand, the eye movement behaviors between different metrics on the same TW can be compared.

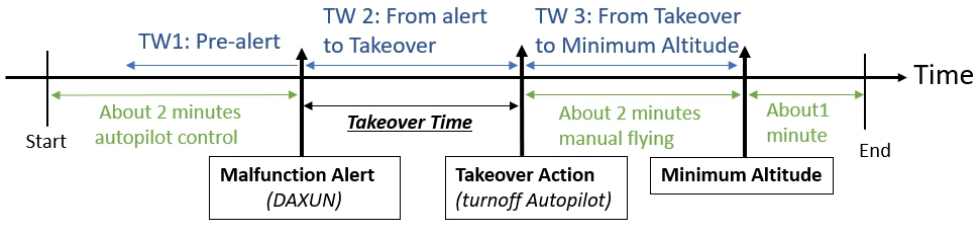
Previous research has demonstrated drivers' behavior changed before and after the hazard was perceived [38]. Therefore, the moment when the hazard scenes were perceived (i.e., the moment when the hazard scene was fixated on) is a critical event during the experimental trial. Following this approach, the current study extracted TW from the whole experiment process using the following two key events: the moment when the TOR

was given and the moment when drivers perceived the hazard scenes. Accordingly, three critical TWs were selected for further analysis: Pre-TOR (TW1), from TOR to hazard perceived (TW2), and from hazard perceived to the end (TW3). Drivers' eye tracking features were analyzed separately within these separate TWs. These three TWs are illustrated in Figure 4.8.

4.3.2 Piloting study



(a) No failure scenarios timeline.



(b) Failure scenarios timeline.

Figure 4.9: Piloting experiment timeline.

In the piloting experiment, the total duration of each scenario varied depending on the power settings and the time taken to detect the malfunction. Given that the entire scenario was relatively long and only certain periods were relevant to pilots' SA and takeover performance, specific TWs were extracted to ensure the analysis focused on the most meaningful segments. On the one hand, the behavior of pilots with better performance can be analyzed across different TWs. On the other hand, eye movement patterns associated with different metrics can be compared within the same TW.

In the piloting experiment, the malfunction alert and the takeover action are two critical events during experiment trial. Following these two events, three critical TWs were extracted for further analysis: Pre-alert (TW1), from alert to Takeover (TW2) and from Takeover to Minimum Altitude (TW3). The "Minimum Altitude" here refers to the lowest

altitude at which the autopilot was engaged, set at 250 feet. These TWs are illustrated in Figure 4.9.

4.4 Data processing

4.4.1 Metrics grouping

To facilitate interpretation and highlight behavioral trends, each SA-related and performance-related metric was split into two groups. Previous literature and domain knowledge were first reviewed to identify theoretically meaningful threshold. However, since the design of experimental scenarios and SAGAT questions were distinct from different experiment, we decided to derive cutoffs from the data itself in the end.

For categorical metrics (i.e. SAGAT score with possible values: 0, 0.5, 1, 1.5, 2), the frequency distribution is used to identify natural groupings (irrelevant or problematic trials excluded). For both drivers' and pilots' SAGAT score results, distribution showed one prominent peak near the score of 1 and another near the score of 2. To create two groups of SAGAT score, the cutoff was placed at 1.5 — the valley between the two peaks. This approach ensures that participants classified as "Higher SAGAT score" (≥ 1.5) answered most SAGAT questions correctly, while those in the "Lower SAGAT score" group (< 1.5) missed a substantial portion, aligning the grouping with meaningful differences in SA.

For continuous metrics such as takeover time, the cutoffs are derived by identifying the point of most pronounced change in the ordered distribution (knee point). Metric values were sorted in ascending order and plotted against their ranking (irrelevant or problematic trials excluded), producing an ordered distribution curve. The cutoff was determined at the knee point of the curve (the point of maximum curvature), where values began to rise disproportionately, indicating a natural separation in the data [5, 16]. This approach, similar to the 'elbow method' used in clustering and scree plot analysis [66], identifies a natural break in the data and ensures that values are grouped based on meaningful behavioral differences.

In all, these methods respects the natural structure of the data while trying to locate a boundary that corresponds to an empirically meaningful transition in behavior or performance. They helps maximize between-group differentiation while reducing within-group heterogeneity, rather than imposing an arbitrary split.

4.4.1.1 Driving study

As introduced in Section 4.2, the drivers' SA is measured by SAGAT score and HPT, and their takeover performance is measured by TOT and Resulting acceleration.

SAGAT score: Figure 4.10 plots the frequency distribution of SAGAT score results from the driving experiment. As introduced above, the cutoff point here is: lower SAGAT score (< 1.5 points) versus higher SAGAT score (≥ 1.5 points).

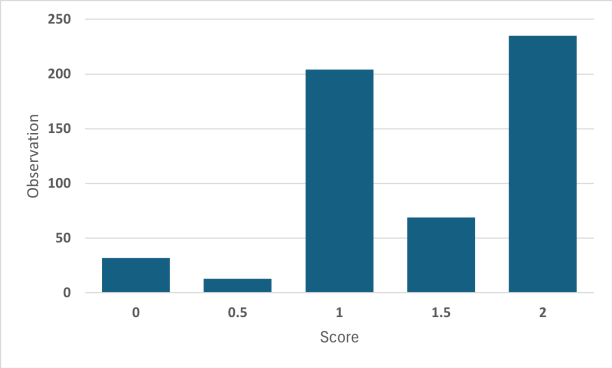


Figure 4.10: SAGAT Score distribution in the driving study

Hazard Perception Time: Figure 4.11 plots the ordered distribution of HPT results from the driving experiment. As introduced above, the cutoff point here is identified by the knee point, which is: longer HPT (≥ 8.35 seconds) versus shorter HPT (< 8.35 seconds).

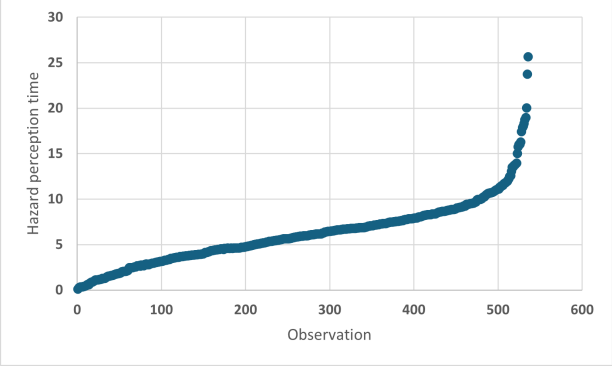


Figure 4.11: Hazard Perception Time distribution in the driving study

Takeover time: Figure 4.12 plots the ordered distribution of TOT results from the driving experiment. As introduced above, the cutoff point here is identified by the knee point, which is: longer TOT ($\geq$ 2.75 seconds) versus shorter TOT ($<$ 2.75 seconds)

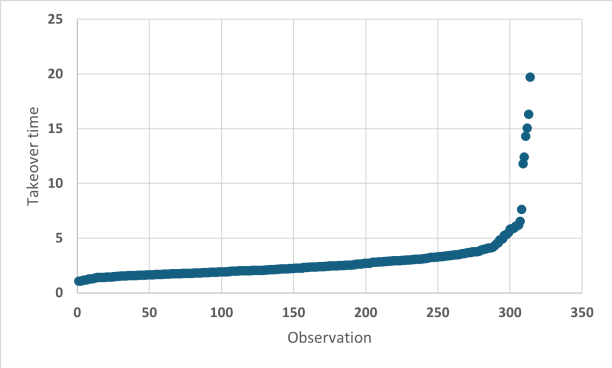


Figure 4.12: Takeover Time distribution in the driving study

Resulting acceleration: Figure 4.13 plots the ordered distribution of Resulting acceleration results from the driving experiment. As introduced above, the cutoff point here is identified by the knee point, which is: higher resulting acceleration ($\geq$ 5 m/s^2) versus lower resulting acceleration ($<$ 5 m/s^2)

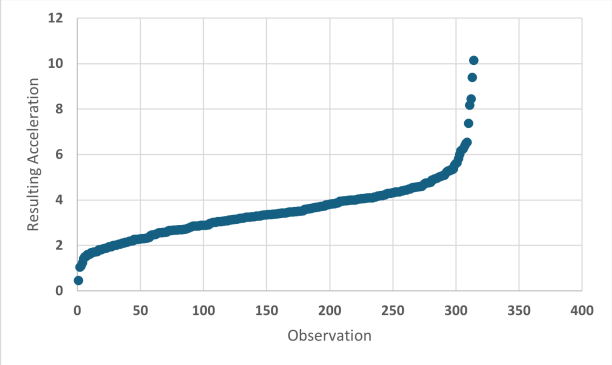


Figure 4.13: Resulting acceleration distribution in the driving study

4.4.1.2 Piloting study

As introduced in Section 4.2, the pilots' SA is measured by SAGAT score and their takeover performance is measured by TOT.

SAGAT score: Figure 4.14 plots the frequency distribution of SAGAT score results from the piloting experiment. As introduced above, the cutoff point here is: lower SAGAT score (< 1.5 points) versus higher SAGAT score (≥ 1.5 points)

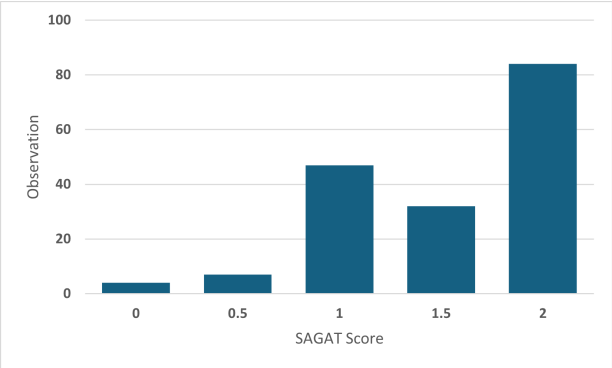


Figure 4.14: SAGAT Score distribution in the piloting study

Takeover time: Figure 4.15 plots the ordered distribution of TOT results from the piloting experiment. As introduced above, the cutoff point here is identified by the knee point, which is: longer TOT (≥ 20 seconds) versus shorter TOT (< 20 seconds)

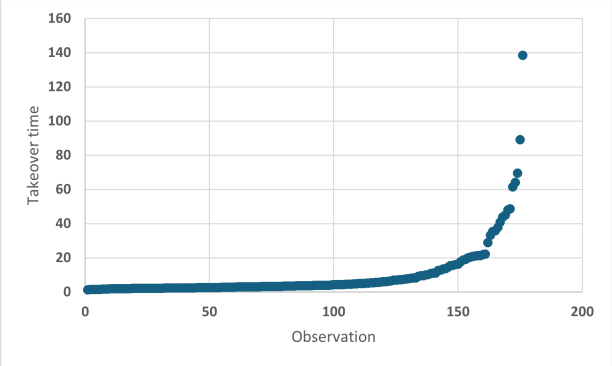


Figure 4.15: Takeover Time distribution in the piloting study

4.4.2 Analytical models

4.4.2.1 MANOVA + Welch's t-tests with FDR

In the Topic 1, we explored the eye movement differences, represented by a series of AOI-based eye tracking features, between SA-related and takeover performance-related metrics. These features include Dwelling time, Entry count, SGE and GTE. And in the Topic 2, the same eye tracking features were investigated between two groups of operators from diverse populations. To investigate above group differences across multiple eye tracking features, a series of MANOVA were conducted using the SciPy library [156] in Python with a significance threshold set as $p = 0.05$. The independent variable is Group (e.g., better- versus worse-SA, or expert vs. learner pilots), and the dependent variables were the set of AOI-based eye tracking features extracted for each trial.

Like introduced before, experimental trials were separated into three distinct TWs by two key events identified during the takeover process. A separate MANOVA was conducted for each TW to evaluate whether group differences varied across different TW. This approach allowed us to treat each window as an independent multivariate comparison, rather than averaging across the entire trial to show different eye movement patterns across different TWs.

MANOVA was chosen instead of running multiple independent t-tests on individual features because it simultaneously evaluates the combined set of dependent variables, accounts for correlations among them, and controls the experiment-wise Type I error rate. Pillai's Trace is reported as the primary multivariate test statistic, given its robustness to violations of the assumption of homogeneity of covariance matrices and multivariate normality [119]. When a statistically significant multivariate effect was identified ($p = 0.05$), follow-up univariate Welch's t-tests were performed on the individual eye tracking features for. These p-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) procedure [17] to control for multiple comparisons and limit the false discovery rate across features.

4.4.2.2 Machine Learning

In Topic 3, machine learning methods were used to predict the operator's Situation Awareness and takeover performance. All machine learning models and data processing techniques were implemented using the scikit-learn package (version 0.21) in Python.

Previous research shows that machine learning models can predict operators' SA and performance effectively. Zhou *et al.* had applied a series of machine learning models

to predict drivers' SA using eye tracking features only, and the result showed that the LightGMB model outperforms other models [172]. Pakdamanian *et al.* proposed a novel neural network-based framework to predict driver's takeover behaviors, and it outperforms previous state-of-the-art machine learning models [121]. These research studies showed that the normal machine learning methods hold the potential to predict operators' SA and takeover performance.

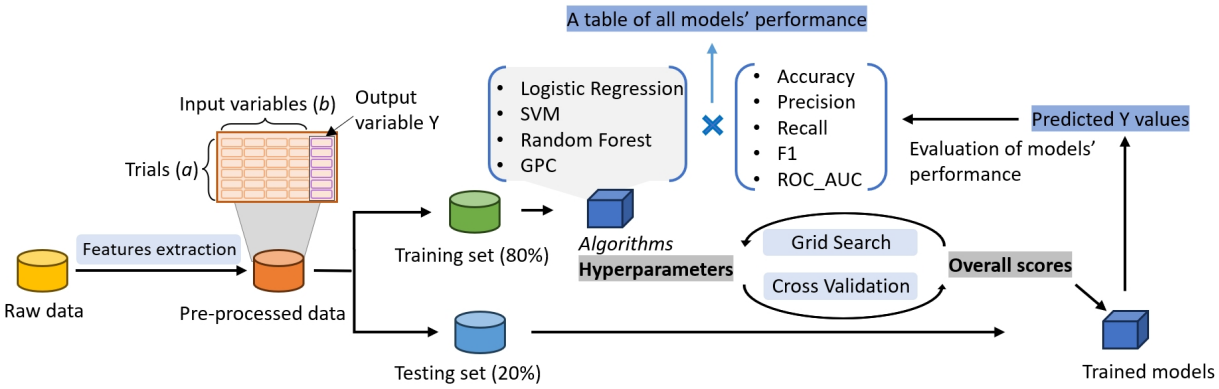


Figure 4.16: Machine Learning pipeline

Figure 4.16 shows the machine learning pipeline used in the current thesis. Data collected in both driving and piloting experiments went through this pipeline with minor differences in the input and output variables. In the pipeline, the raw data includes the eye tracking data as input, SA-related metrics and takeover performance-related metrics as output. The pre-processed dataset was obtained from the eye tracking features extraction, which had been discussed in the previous Section 4.1.

The dataset is in a $a \times (b+1)$ structure, where a is the total number of trials conducted in the experiment and b is the total number of eye tracking features. The additional column $b+1$ represents the target metric to be predicted. The pre-processed data was split into 80% as the training set and 20% as the testing set. Four supervised classification models were trained on the training set: Logistic Regression, Support Vector Machine (SVM), Random Forest (RF) and Gaussian Process Classifier (GPC). Model performance was evaluated using five standard evaluation criteria for classification: Accuracy, Precision, Recall, F1 Score, and AUC.

The hyperparameters of models used here were also optimized, such as the kernel in SVM and the number of trees in RF. Considering that not too many hyperparameters were included in the models chosen above, running a few iterations of the models would be

tedious. Therefore, we used grid search with the k-fold cross validation ($k = 5$) method to find the best hyperparameters for the final trained models. The training set was divided into five folds, and each fold was used as the validation set for one time when the other four folds were used to train the model. An overall performance score was obtained by averaging the evaluation results over five iterations. All possible combinations of hyperparameters went through this process, and their overall score yielded the best hyperparameter combination. Thus, final trained models were obtained. Before model training, all features were standardized using a StandardScaler to ensure that each had zero mean and unit variance. To reduce dimensionality and retain the most informative features, PCA was applied, preserving 95% of the original variance. These steps were integrated prior to grid search and cross-validation. If we fed the testing set into a trained model, a predicted Y value can be obtained. The predicted Y value was compared with the true Y value in the testing set by five different evaluation approaches listed above.

Although the most important hyperparameters were optimized here, some parameters are still predetermined, such as k in the cross-validation and evaluation metrics in cross-validation. It was too complicated and tedious to optimize everything in one pipeline. Those parameters were optimized later by rerunning this machine learning pipeline with new parameter values and comparing the outcomes with the current ones.

Chapter 5

Differences eye movement between drivers with different Situation Awareness levels and takeover performance levels in a semi-autonomous driving system.

This chapter presents the driving domain results and discussion of Topic 1: **Differences in eye movement between operators with different Situation Awareness levels and takeover performance levels in a conditionally automated system.**

5.1 Background

Under the conditions of rapid development of semi-autonomous vehicles, human drivers are facing benefits as well as challenges at the same time. On one hand, the autonomous driving system can provide transportation without human intervention within its capacity, therefore relieving human beings from tedious driving tasks. On the other hand, level 3 autonomous driving systems have their own ODD. According to the latest taxonomy from SAE released, level 3 autonomous cars expect human drivers to take control of the vehicles back when facing occasions recognized as unmanageable by the control system, known as the takeover process.

The takeover process could be challenging for drivers, primarily due to their limited SA of the surrounding environment after takeover, especially if they have been engaged in NDRTs during the automation period [33]. When suddenly required to regain control, drivers may experience inadequate SA as they switch their attention back to the driving task. This transition period can lead to decreased takeover performance, such as delayed reaction times, imprecise steering inputs, and unstable vehicle control.

Eye tracking has emerged as a powerful tool to investigate these issues, offering detailed insights into drivers' SA and takeover performance. Understanding drivers' eye movement differences between drivers with different SA levels and takeover performance levels during the takeover process in semi-autonomous vehicles is crucial for making explainable conclusions to ensure safety. The current study focuses on four types of eye-tracking AOI features, including Dwelling time, Entry count, SGE and GTE to explore how drivers with different SA and takeover performance levels behaved differently on the features above.

5.2 Methodology

One simulated takeover driving experiment was conducted to investigate the research question: How does drivers with different SA levels and takeover performance levels move their eyes differently in a semi-autonomous driving system. The details of the driving experiment is covered in Chapter 3.1.

Drivers' eye movement was collected by eye tracking glasses and their AOI-based eye tracking features were extracted following the steps introduced in Chapter 4.1.3.1. Four categories of AOI features were used to represent drivers' eye movement (see Table 4.2): Dwelling time, Entry Count, SGE and GTE. Since each type of feature represents a different aspect of drivers' eye movement, they were discussed one by one in the results section. After processing the data, we found that the dwelling time and entry counts of three mirrors: Rear-mirror, Left mirror, and Right mirror were relatively low. Therefore, these three AOIs were combined into one single AOI labeled as the "Mirrors".

To draw insights from the analysis, eye tracking features were compared between different TWs (defined in Chapter 4.3.1) to examine drivers' eye movement differences between groups as well as changes through time. Consistent with the overall analytical framework introduced in Chapter 4.4.2.1, MANOVA was used to test for multivariate effects, followed by Welch's t-tests for pairwise comparisons when appropriate.

In the data analysis process, some revisions were made towards TW separation. Previous timeline defined three TWs used for the driving experiment (see Figure 4.8), and the

first TW was "Pre-TOR". After testing time periods of "15 seconds", "10 seconds" and "5 seconds", it was found that "10 seconds" can best represent the behavior of drivers before the TOR. Therefore, this TW was finally set to be "10 seconds before the TOR" in the current driving experiment.

In all, each driving experimental trial was separated into three TWs, and all AOI features were extracted separately for each of the three TWs. As described in Chapter 3.1, the driving scenarios included manual, automation without NDRT and automation with NDRT, where the NDRT included two visual tasks and one auditory task. However, not all trials were suitable for analysis. For example, since the AOI setups in the driving experiment assumed drivers were viewing the front monitor, trials involving visual distractions were excluded. Because these trials required drivers to engage with a separate side laptop, which meant they were not looking at the front screen where the AOIs were defined. In the end, for the investigation of SA, all remaining experimental trials were included. This included manual trials, as they still allowed examination of how drivers used eye movements to gather situational information across the three TWs, thereby enabling comparisons between different SA groups. In contrast, for the analysis of takeover performance, all manual trials were excluded, as they did not involve any takeover event by design.

Additionally, we examined whether eye movement features exhibit distinct patterns across two metrics representing the same underlying construct. For example, if Dwelling time in an AOI shows distinct patterns between drivers with lower versus higher SAGAT scores, and also between those with longer versus shorter HPT, it suggests that these two metrics are not easily optimized simultaneously.

5.3 Results

5.3.1 Drivers' Situation Awareness

Drivers' SA was measured using two metrics: the SAGAT score and HPT. For each of the three TWs, separate MANOVA tests using Pillai's Trace were conducted twice — once for the two SAGAT score groups and once for the two HPT groups (the grouping criteria were introduced in Chapter 4.4.1). These analyses examined whether the overall eye movement patterns differed significantly between groups within each TW. When a significant multivariate effect was observed, follow-up Welch's independent t-tests were performed to determine which specific AOI-based features contributed to the observed differences.

Results for higher versus lower SAGAT score groups:

- TW1: Pillai's Trace = 0.010, $F(11, 467) = 0.431, p = 0.942$. No significant multivariate effect was observed, suggesting that eye movement patterns did not differ between higher and lower SAGAT score groups during this time window.
- TW2: Pillai's Trace = 0.045, $F(11, 467) = 2.010, p = 0.026$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between higher and lower SAGAT score groups.
- TW3: Pillai's Trace = 0.041, $F(11, 467) = 1.810, p = 0.049$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between higher and lower SAGAT score groups

Results for shorter versus longer HPT groups:

- TW1: Pillai's Trace = 0.022, $F(11, 467) = 0.967, p = 0.476$. No significant multivariate effect was observed, suggesting that eye movement patterns did not differ between shorter and longer HPT groups during this time window.
- TW2: Pillai's Trace = 0.153, $F(11, 467) = 7.665, p = 0.000$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between shorter and longer HPT groups.
- TW3: Pillai's Trace = 0.040, $F(11, 467) = 1.780, p = 0.051$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between shorter and longer HPT groups.

To facilitate interpretation, the follow-up Welch's t-test results are presented in the subsequent sections, organized by feature category (Dwelling Time, Entry Count, and Entropy). Each section reports findings across all three TWs and summarizes the key AOI specific results. Although the multivariate effects in the TW1 were not significant for both SA metrics, we still report the Welch's t-test results below to keep the consistency.

5.3.1.1 Dwelling time

Drivers with different levels of SA might exhibit different Dwelling time features on various AOIs across different TWs. The dwelling time features contain "Mirrors_DT", "Center_DT", "Dashboard_DT" and "Others_DT", corresponding to the 6 AOIs defined based

on experiment scenarios, with three mirrors combined into a single AOI named "Mirrors". Each feature represents the percentage of time that drivers spent in the AOI.

Table 5.1 shows results of the Welch's independent t-tests; the left section represents the result of SAGAT score, and the right section represents the result of HPT. Figure 5.1 demonstrates the Dwelling time of all AOIs across three TWs. Each pair of bars represents the values from two different groups of drivers. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this Dwelling time feature between the two groups. The error bars represent SEM of each bar. However, although Welch's t-tests were conducted for TW1 for consistency, their results are not interpreted further because the multivariate effect was not significant.

Figure 5.1a, Figure 5.1c and Figure 5.1e demonstrate the results for two SAGAT score groups across three different TWs, displayed from top to bottom. Figure 5.1b, Figure 5.1d and Figure 5.1f demonstrate the results for two HPT groups across three different TWs, displayed from top to bottom. The two figures in each row correspond to the same TW, allowing a direct comparison between eye movement patterns between drivers from the two metric groups.

Table 5.1: T-tests for Dwelling times features between drivers from two SAGAT score groups and two HPT groups across three TWs (* indicates $p < .05$)

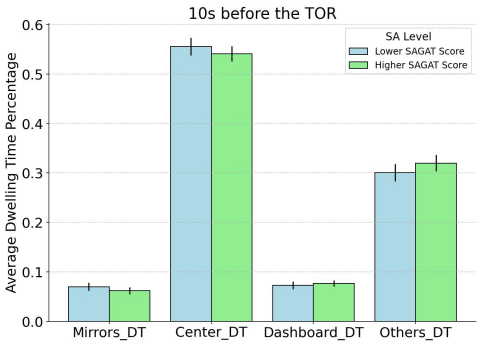
	Lower vs. Higher SAGAT Score				**Longer vs. Shorter Hazard Perception Time**			
	Mirrors	**Center**	**Dashboard**	**Others**	**Mirrors**	**Center**	**Dashboard**	**Others**
TW1								
T-value	−0.783	−0.622	0.371	0.786	0.113	−0.851	−0.554	1.013
P-value	0.434	0.534	0.710	0.432	0.910	0.395	0.580	0.312
Effect size	0.072	0.057	0.034	0.072	0.013	0.093	0.580	0.113
TW2								
T-value	−0.433	2.197	1.665	−2.684	0.319	2.848	0.194	−2.854
P-value	0.665	**0.028***	0.097	**0.008***	0.750	**0.005***	0.846	**0.005***
Effect size	0.039	0.201	0.153	0.247	0.039	0.314	0.021	0.309
TW3								
T-value	0.875	−1.502	−1.124	1.585	0.391	2.304	0.237	−2.466
P-value	0.382	0.134	0.261	0.114	0.696	**0.022***	0.813	**0.014***
Effect size	0.080	0.138	0.101	0.147	0.043	0.247	0.026	0.264

Note: The "DT" in feature names is omitted to save space.

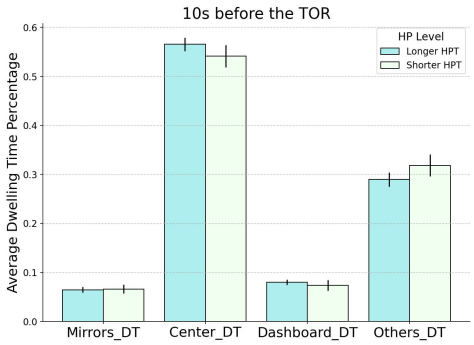
Results

- SAGAT scores: From TOR to hazard perceived (TW2; Figure 5.1c), drivers with higher SAGAT scores exhibited significantly longer Dwelling time in "Center" area, while significantly shorter Dwelling time in "Other" area compared to those with lower SAGAT scores. This suggests that drivers with higher SAGAT scores directed their attention more toward the central field of view, potentially to monitor critical road elements. After the hazard was perceived (TW3; Figure 5.1e), this trend disappeared and no significance was shown on any AOI.

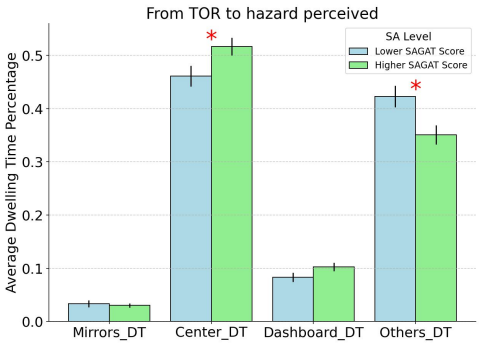
- Hazard perception time: In both TWs—TOR to hazard perceived (TW2; Figure 5.1d), and hazard perceived till the end (TW3; Figure 5.1f)—drivers with shorter HPT consistently spent more time focusing on "Center" area and less time on "Others" area. This result suggests that drivers with shorter HPT tended to allocate more visual attention to the center of the street and less to the peripheral areas during TW2, likely to detect the hazard quickly, as it typically appears in the center of the road. Interestingly, this focused attention persisted even after hazard detection, indicating a faster hazard recognition may also lead to a stable attentional pattern.



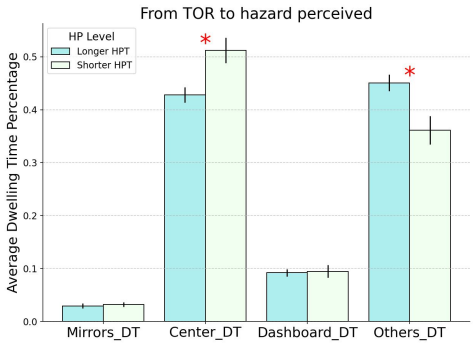
(a) SAGAT Score: TW1



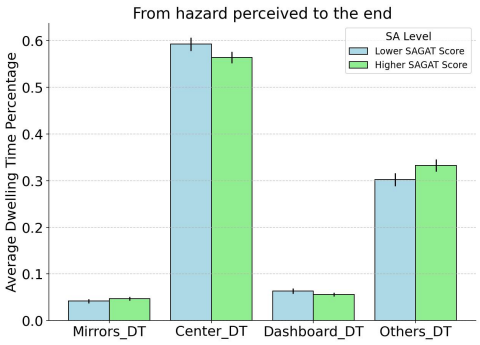
(b) Hazard Perception Time: TW1



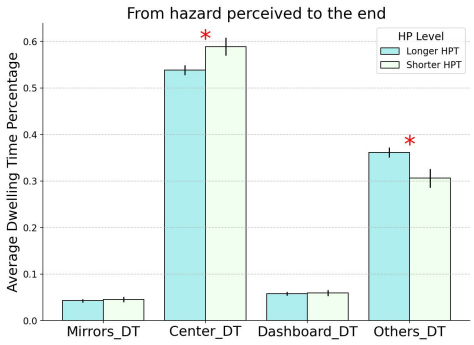
(c) SAGAT Score: TW2



(d) Hazard Perception Time: TW2



(e) SAGAT Score: TW3



(f) Hazard Perception Time: TW3

Figure 5.1: Dwelling time features of drivers from two SAGAT score groups and two HPT groups across three TWs (error bars represent SEM).

5.3.1.2 Entry Count

Drivers with different levels of SA might exhibit different Entry count features on various AOIs across different TWs. The Entry count features contain "Mirrors_EC", "Center_EC", "Dashboard_EC", "Screen_EC" and "Others_EC". Each feature represents the number of times that drivers entered into that AOI.

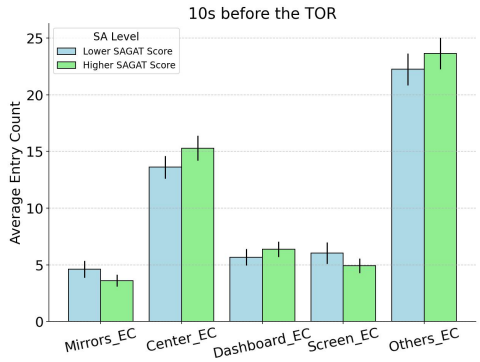
Table 5.2: T-tests for Entry count features between drivers from two SAGAT score groups and two HPT groups across three TWs (* indicates $p < .05$)

	Lower vs. Higher SAGAT Score					**Longer vs. Shorter Hazard Perception Time**				
	Mirrors	**Center**	**Dashboard**	**Screen**	**Others**	**Mirrors**	**Center**	**Dashboard**	**Screen**	**Others**
TW1										
T-value	−1.153	1.104	0.702	−1.007	0.709	−0.903	−0.129	−0.081	−1.966	−0.433
P-value	0.249	0.269	0.483	0.315	0.479	0.367	0.897	0.936	**0.050***	0.665
Effect size	0.105	0.103	0.064	0.091	0.066	0.096	0.015	0.009	0.199	0.048
TW2										
T-value	−1.788	1.360	1.334	−1.041	0.859	−1.601	−6.450	−3.058	−6.387	−8.252
P-value	0.074	0.174	0.183	0.298	0.391	0.110	**0.000***	**0.002***	**0.000***	**0.000***
Effect size	0.157	0.128	0.124	0.094	0.079	0.182	0.579	0.580	0.316	0.761
TW3										
T-value	0.571	2.795	−0.521	1.121	2.378	1.421	0.527	−1.267	−0.038	0.872
P-value	0.568	**0.005***	0.602	0.263	**0.018***	0.156	0.599	0.206	0.969	0.384
Effect size	0.051	0.261	0.047	0.105	0.221	0.178	0.059	0.004	0.132	0.098

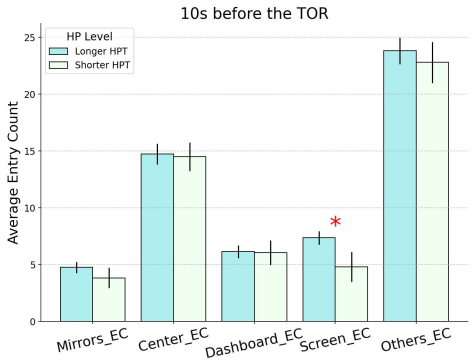
Note: The "EC" in feature names is omitted to save space. "Screen" refers to the extra Entry count AOI feature.

Table 5.2 shows the results of Welch's independent t-tests; the left section represents the result of SAGAT score, and the right section represents the result of HPT. Figure 5.2 demonstrates the Entry count of all AOIs across three TWs. Each pair of bars represents the values from two different groups of drivers. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this Entry count feature between the two groups. The error bars represent SEM of each bar. However, although Welch's t-tests were conducted for TW1 for consistency, their results are not interpreted further because the multivariate effect was not significant.

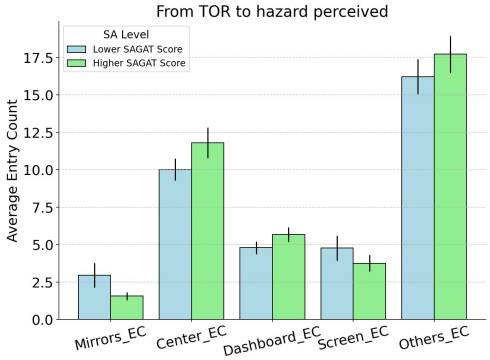
Figure 5.2a, Figure 5.2c and Figure 5.2e demonstrate the results from two SAGAT score groups across three different TWs and Figure 5.2b, Figure 5.2d and Figure 5.2f demonstrate the results from two HPT groups on three different TWs.



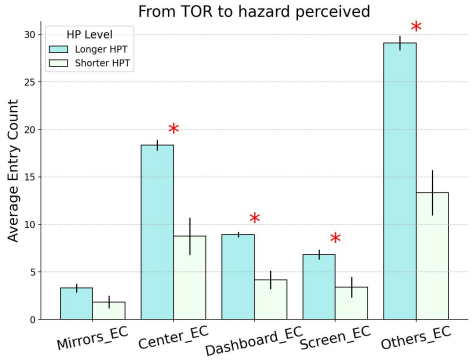
(a) SAGAT Score: TW1



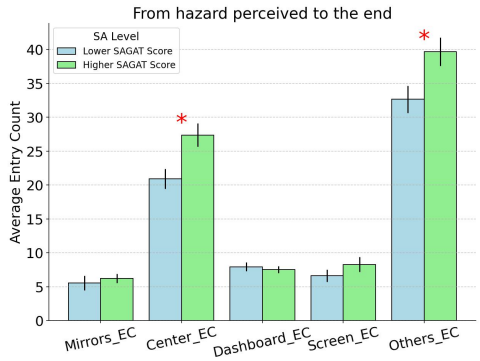
(b) Hazard Perception Time: TW1



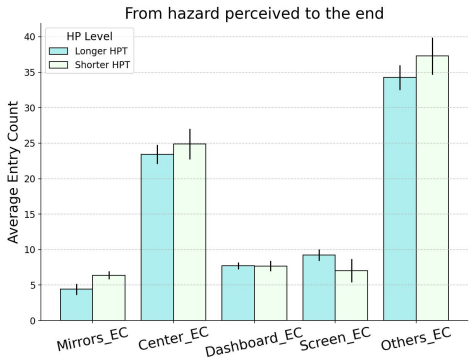
(c) SAGAT Score: TW2



(d) Hazard Perception Time: TW2



(e) SAGAT Score: TW3



(f) Hazard Perception Time: TW3

Figure 5.2: Entry count features of drivers from two SAGAT score groups and two HPT groups across three TWs (error bars represent SEM).

Results

- SAGAT scores: From hazard perceived till the end of the trial (TW3; Figure 5.2e), after the hazard had already been perceived and manual control resumed, drivers with higher SAGAT exhibited more Entry counts in "Center" and "Others". This result showed that the drivers with higher SAGAT score continued actively scanning the environment to gather situational information. This happened after they had already perceived the hazards and taken over manual control of the vehicle. No significant differences were found in the TW2.

- Hazard perception time: From TOR to hazard perceived (TW2; Figure 5.2d), drivers with shorter HPT exhibited less Entry counts in all AOIs except the "Mirrors". This result suggests that drivers who perceived the hazard scenes with shorter time entered into different AOIs with fewer times. They showed more assured eye movement, likely allowing them to detect the hazard more efficiently. No significant differences or interpretable results were found in TW3.

5.3.1.3 Entropies

Drivers with different levels of SA might exhibit different SGE and GTE on various AOIs across different TWs. The calculation of these two entropy values was introduced in Chapter 4.1.3.

Table 5.3 shows the results of Welch's independent t-tests; the left section represents the result of SAGAT score, and the right section represents the result of HPT. Figure 5.3 demonstrates the SGE and GTE of all AOIs across three TWs. Each pair of bars represents the values from two different groups of drivers. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this entropy feature between the two groups. The error bars represent SEM of each bar. However, although Welch's t-tests were conducted for TW1 for consistency, their results are not interpreted further because the multivariate effect was not significant.

Figure 5.3a, Figure 5.3c and Figure 5.3e demonstrate the results from two SAGAT score groups across three different TWs and Figure 5.3b, Figure 5.3d and Figure 5.3f demonstrate the results from two HPT groups on three different TWs.

Results

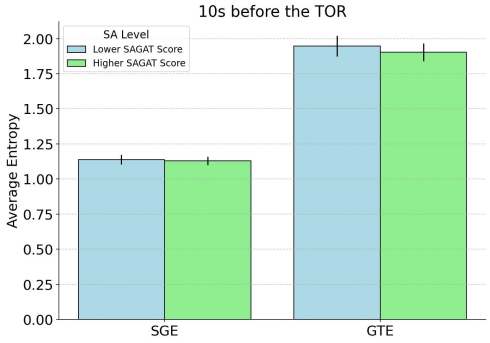
Table 5.3: T-tests for SGE and GTE between drivers from two SAGAT score groups and two HPT groups across three TWs (* indicates $p < .05$)

	Lower vs. Higher SAGAT Score		**Longer vs. Shorter Hazard Perception Time**	
	SGE	**GTE**	**SGE**	**GTE**
TW1				
T-value	−0.220	−0.452	−1.001	−2.055
P-value	0.826	0.652	0.317	**0.040***
Effect size	0.020	0.041	0.109	0.218
TW2				
T-value	0.303	1.142	−1.941	−4.677
P-value	0.762	0.254	0.053	**0.000***
Effect size	0.028	0.104	0.215	0.525
TW3				
T-value	1.316	2.005	−0.758	0.442
P-value	0.189	**0.045***	0.449	0.659
Effect size	0.121	0.184	0.079	0.048

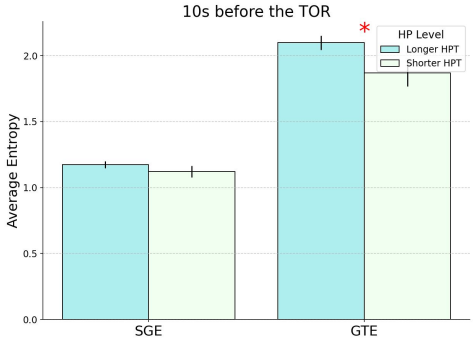
Note: SGE = Stationary Gaze Entropy; GTE = Gaze Transition Entropy.

- SAGAT scores: From hazard perceived till the end of the trial (TW3; Figure 5.3e), drivers with higher SAGAT scores exhibited a significantly higher GTE compared the drivers with lower SAGAT scores. This result suggests that they had a more complex and uncertain scanning pattern between different areas, which shows some degree of consistency the Entry count result in the same TW where drivers with higher SAGAT scores entered two AOIs significantly more times.

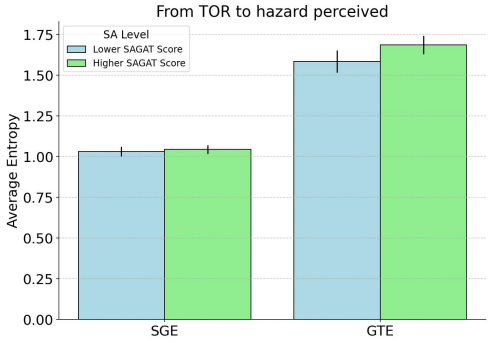
- Hazard perception time: From TOR to hazard perceived (TW2; Figure 5.3d), drivers with shorter HPT showed significantly lower GTE than drivers with longer HPT. This result also shows some consistency with the Entry count results in TW2 where shorter-HPT drivers had significantly fewer entries into various AOIs. During this time period, shorter-HPT drivers exhibited a more focused visual movement with less uncertainty, possibly due to they located the target areas faster than longer-HPT drivers.



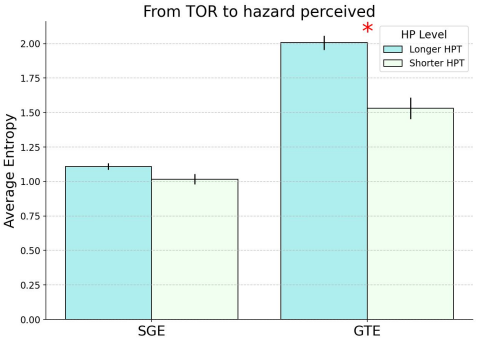
(a) SAGAT Score: TW1



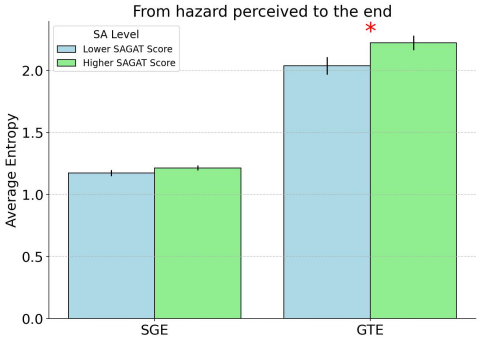
(b) Hazard Perception Time: TW1



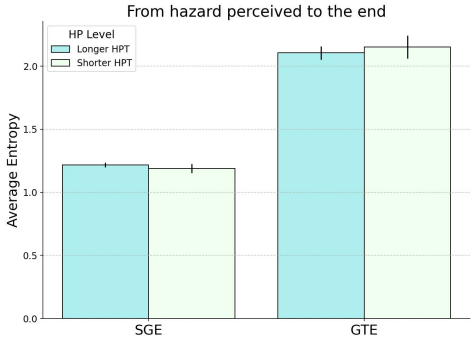
(c) SAGAT Score: TW2



(d) Hazard Perception Time: TW2



(e) SAGAT Score: TW3



(f) Hazard Perception Time: TW3

Figure 5.3: SGE and GTE of drivers from two SAGAT score groups and two HPT groups across three TWs (error bars represent SEM).

5.3.2 Drivers' Takeover performance

Drivers' Takeover performance was measured using two metrics: TOT and Resulting acceleration. For each of the three TWs, separate MANOVA tests using Pillai's Trace were conducted twice — once for the two TOT groups and once for the two Resulting acceleration groups (the grouping criteria were introduced in Chapter 4.4.1). These analyses examined whether the overall eye movement patterns differed significantly between groups within each TW. When a significant multivariate effect was observed, follow-up Welch's independent t-tests were performed to determine which specific AOI-based features contributed to the observed differences.

Results for shorter versus longer TOT groups:

- TW1: Pillai's Trace = 0.096, $F(11, 467) = 4.515, p = 0.000$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between shorter versus longer TOT groups.
- TW2: Pillai's Trace = 0.067, $F(11, 467) = 3.045, p = 0.001$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between shorter versus longer TOT groups.
- TW3: Pillai's Trace = 0.042, $F(11, 467) = 1.877, p = 0.040$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between shorter versus longer TOT groups.

Results for lower versus higher Resulting acceleration groups:

- TW1: Pillai's Trace = 0.064, $F(11, 467) = 2.917, p = 0.001$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between lower versus higher Resulting acceleration groups.
- TW2: Pillai's Trace = 0.012, $F(11, 467) = 0.515, p = 0.894$. No significant multivariate effect was observed, suggesting that eye movement patterns did not differ between lower versus higher Resulting acceleration groups.
- TW3: Pillai's Trace = 0.049, $F(11, 467) = 2.178, p = 0.015$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between lower versus higher Resulting acceleration groups.

To facilitate interpretation, the follow-up Welch's t-test results are presented in the subsequent sections, organized by feature category (Dwelling Time, Entry Count, and Entropy). Each section reports findings across all three TWs and summarizes the key AOI specific results. Although the multivariate effects in the TW2 were not significant for Resulting acceleration, we still report the Welch's t-test results below to keep the consistency.

5.3.2.1 Dwelling time

Drivers with different levels of Takeover performance might exhibit different Dwelling time features on various AOIs across different TWs. The dwelling time features contain "Mirrors_DT", "Center_DT", "Dashboard_DT" and "Others_DT", corresponding to the 6 AOIs defined based on experiment scenarios, with three mirrors combined into a single AOI named "Mirrors". Each feature represents the percentage of time that drivers spent in the AOI.

Table 5.4: T-tests for Dwelling Time features between drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (* indicates $p < .05$)

	Longer vs. Shorter Takeover Time				**Higher vs. Lower Resulting Acceleration**			
	Mirrors	**Center**	**Dashboard**	**Others**	**Mirrors**	**Center**	**Dashboard**	**Others**
TW1								
T-value	−0.672	−1.340	−1.857	2.383	2.456	−1.543	−0.987	0.695
P-value	0.502	0.182	0.065	**0.018***	**0.015***	0.124	0.325	0.488
Effect size	0.094	0.182	0.272	0.319	0.386	0.344	0.294	0.160
TW2								
T-value	−1.015	3.678	0.845	−3.417	0.811	−1.138	−1.735	1.579
P-value	0.311	**0.000***	0.399	**0.001***	0.418	0.256	0.084	0.116
Effect size	0.145	0.499	0.113	0.463	0.144	0.251	0.492	0.358
TW3								
T-value	−0.638	−0.318	−1.162	0.943	0.304	−2.646	0.809	2.121
P-value	0.524	0.751	0.246	0.347	0.762	**0.009***	0.419	**0.035***
Effect size	0.088	0.043	0.160	0.126	0.054	0.632	0.174	0.502

Note: The "DT" in feature names is omitted to save space.

Table 5.4 shows the results of Welch's t-tests; the left section represents the result of TOT, and the right section represents the result of Resulting acceleration. Figure 5.4

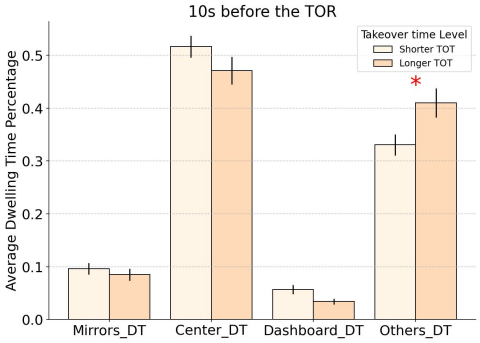
demonstrates the Dwelling time of all AOIs across three TWs. Each pair of bars represents the values from two different groups of drivers. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this Dwelling time feature between the two groups. The error bars represent SEM of each bar. Although Welch's t-tests were conducted for Resulting Acceleration during TW2 for consistency, their results are not interpreted further because the multivariate effect was not significant.

Figure 5.4a, Figure 5.4c and Figure 5.4e demonstrate the results for two TOT groups across three different TWs, displayed from top to bottom. Figure 5.4b, Figure 5.4d and Figure 5.4f demonstrate the results for two Resulting acceleration groups across three different TWs, displayed from top to bottom. The two figures in each row correspond to the same TW, allowing a direct comparison between eye movement patterns between drivers from the two metric groups.

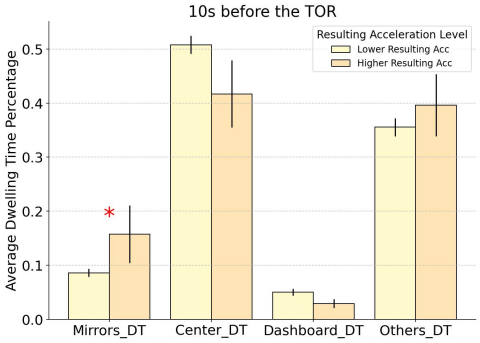
Results

- Takeover Time: The results showed that drivers with shorter TOT spent significantly less time focused at the "Others" area during the 10 seconds before the TOR (TW1; Figure 5.4a). However, this trend reversed after the TOR issued. From the TOR to the moment the hazard was perceived (TW2; Figure 5.4c), shorter-TOT drivers spent significantly less time in the "Center" area and significantly more time in the "Others" area. This suggests that before the TOR, shorter-TOT drivers focused more on information-critical areas — such as the "Mirrors," "Center," and "Dashboard." After TOR issued, their attention shifted toward the peripheral areas, possibly to gather more diverse environmental information, as the "Others" area includes roadside vehicles and pedestrians.

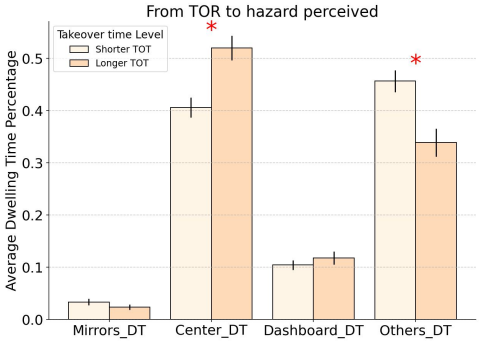
- Resulting Acceleration: Drivers with lower resulting acceleration spent significantly more time focusing on the "Center" area and significantly less time on the "Others" area after the hazard was perceived (TW3; Figure 5.4f). This may indicate that these drivers maintained their gaze on the center of the road to support more stable vehicle control.



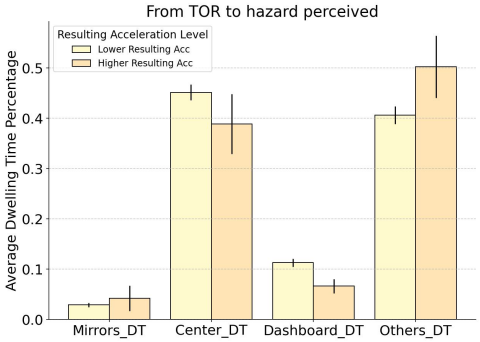
(a) Takeover Time: TW1



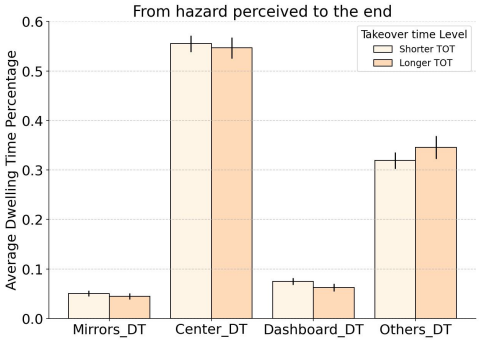
(b) Resulting Acceleration: TW1



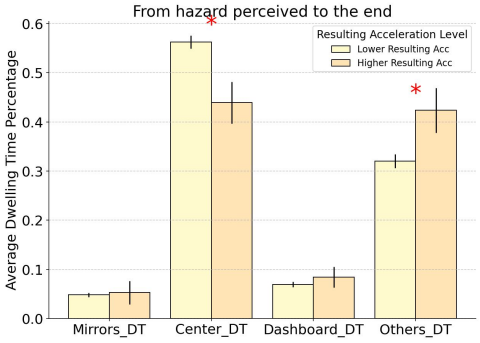
(c) Takeover Time: TW2



(d) Resulting Acceleration: TW2



(e) Takeover Time: TW3



(f) Resulting Acceleration: TW3

Figure 5.4: Dwelling time features of drivers from two Takeover time groups and two Resulting Acceleration groups across three TWs (error bars represent SEM).

5.3.2.2 Entry count

Drivers with different levels of Takeover performance might exhibit different Entry count features on various AOIs across different TWs. The Entry count features contain "Mirrors_EC", "Center_EC", "Dashboard_EC", "Screen_EC" and "Others_EC". Each feature represents the number of times that drivers entered into that AOI.

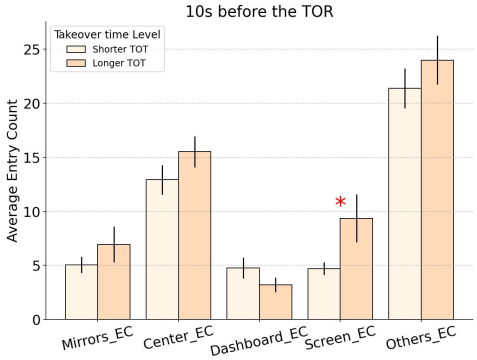
Table 5.5 shows the results of Welch's independent t-tests; the left section represents the result of TOT, and the right section represents the result of Resulting acceleration. Figure 5.5 demonstrates the Entry count of all AOIs across three TWs. Each pair of bars represents the values from two different groups of drivers. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this Entry count feature between the two groups. The error bars represent SEM of each bar. Although Welch's t-tests were conducted for Resulting Acceleration during TW2 for consistency, their results are not interpreted further because the multivariate effect was not significant.

Figure 5.5a, Figure 5.5c and Figure 5.5e demonstrate the results from two TOT groups across three different TWs and Figure 5.5b, Figure 5.5d and Figure 5.5f demonstrate the results from two Resulting acceleration groups on three different TWs.

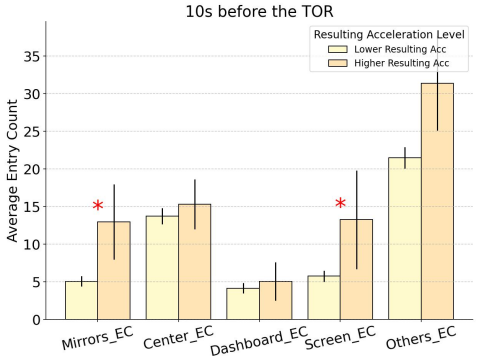
Table 5.5: T-tests for Entry count features between drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (* indicates $p < .05$)

	Longer vs. Shorter Takeover time					**Higher vs. Lower Resulting Acceleration**				
	Mirrors	**Center**	**Dashboard**	**Screen**	**Others**	**Mirrors**	**Center**	**Dashboard**	**Screen**	**Others**
TW1										
T-value	1.194	1.216	−1.128	2.548	0.872	2.916	0.428	0.377	2.363	1.920
P-value	0.234	0.225	0.260	**0.011***	0.384	**0.004***	0.669	0.707	**0.019***	0.056
Effect size	0.149	0.169	0.164	0.301	0.119	0.453	0.102	0.084	0.337	0.395
TW2										
T-value	−0.520	2.095	0.728	1.299	1.319	0.951	0.198	−0.430	1.052	0.264
P-value	0.603	**0.037***	0.468	0.195	0.188	0.343	0.843	0.667	0.294	0.792
Effect size	0.074	0.259	0.099	0.162	0.171	0.166	0.049	0.127	0.182	0.064
TW3										
T-value	0.331	4.177	−0.455	1.925	3.279	2.537	−0.293	0.359	0.046	0.998
P-value	0.741	**0.000***	0.649	0.055	**0.001***	**0.012***	0.770	0.720	0.963	0.319
Effect size	0.047	0.526	0.060	0.234	0.422	0.289	0.079	0.087	0.014	0.222

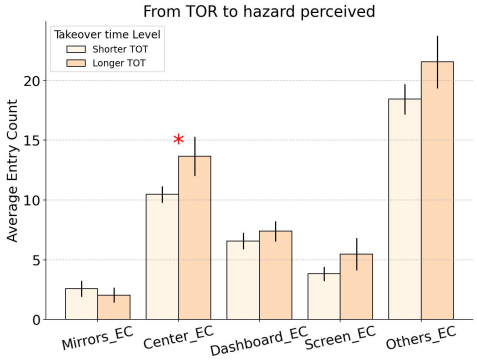
Note: The "EC" in feature names is omitted to save space. "Screen" refers to the extra Entry count AOI feature.



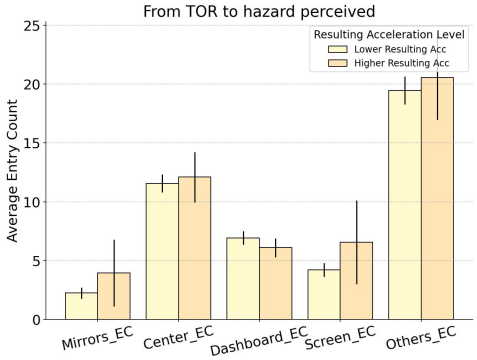
(a) Takeover Time: TW1



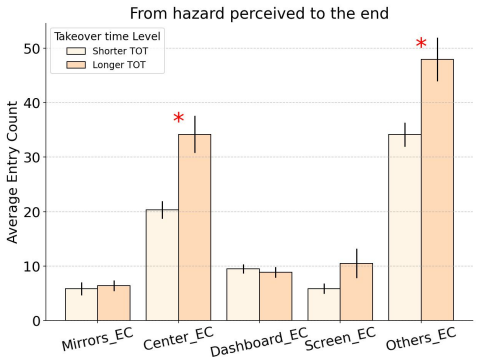
(b) Resulting Acceleration: TW1



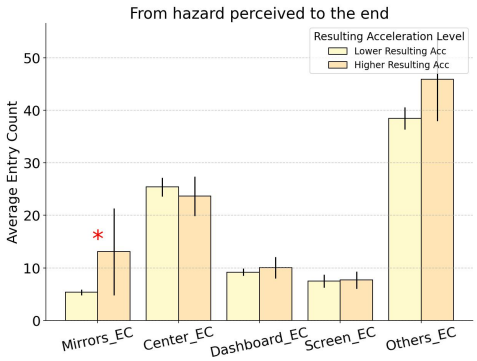
(c) Takeover Time: TW2



(d) Resulting Acceleration: TW2



(e) Takeover Time: TW3



(f) Resulting Acceleration: TW3

Figure 5.5: Entry count features of drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (error bars represent SEM).

Results

- Takeover Time: During the 10 seconds before the TOR (TW1; Figure 5.5a), drivers with shorter TOT made significantly fewer entries into the "Screen" area, which encompasses several AOIs. From TOR to hazard perceived (TW2; Figure 5.5c), drivers with shorter TOT had significantly fewer entries into the "Center" compared to those with longer TOT. From hazard perception to the end (TW3; Figure 5.5e), these drivers also showed significantly fewer entries into both the "Center" and "Others" AOIs.

- Resulting Acceleration: During the 10 seconds before the TOR (TW1; Figure 5.5b), drivers with lower Resulting acceleration showed fewer entries into both the "Mirrors" and "Screen" areas. From hazard perception to the end (TW3; Figure 5.5f), drivers with lower resulting acceleration made significantly fewer entries into the "Mirrors" compared to those with higher resulting acceleration.

5.3.2.3 Entropy

Drivers with different levels of Takeover performance might exhibit different SGE and GTE across different TWs.

Table 5.6 shows the results of Welch's independent t-tests; the left section represents the result of TOT, and the right section represents the result of Resulting acceleration. Figure 5.6 demonstrates the Entry count of all AOIs across three TWs. Each pair of bars represents the values from two different groups of drivers. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this Entry count feature between the two groups. Although Welch's t-tests were conducted for Resulting Acceleration during TW2 for consistency, their results are not interpreted further because the multivariate effect was not significant.

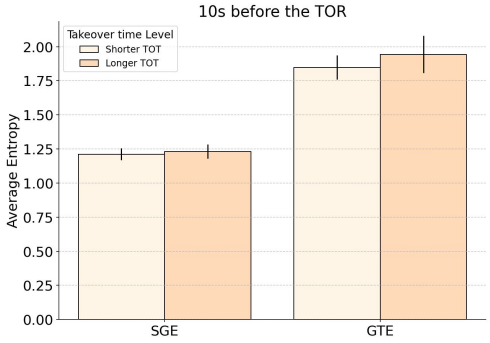
Figure 5.6a, Figure 5.6c and Figure 5.6e demonstrate the results from two TOT groups across three different TWs and Figure 5.6b, Figure 5.6d and Figure 5.6f demonstrate the results from two Resulting acceleration groups on three different TWs.

Table 5.6: T-tests for SGE and GTE between drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (* indicates $p < .05$)

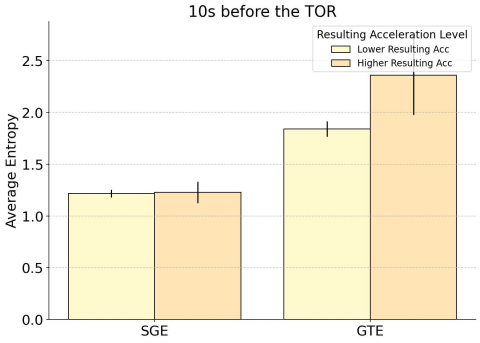
	Longer vs. Shorter Takeover Time		**Higher vs. Lower Resulting Acceleration**	
	SGE	**GTE**	**SGE**	**GTE**
TW1				
T-value	0.274	0.611	0.095	1.921
P-value	0.784	0.542	0.925	0.056
Effect size	0.037	0.081	0.023	0.362
TW2				
T-value	0.291	2.630	−1.174	0.425
P-value	0.771	**0.009***	0.242	0.672
P-value	0.040	0.363	0.316	0.102
TW3				
T-value	−0.411	1.903	0.704	−0.143
P-value	0.682	0.058	0.482	0.886
P-value	0.054	0.249	0.174	0.034

Results

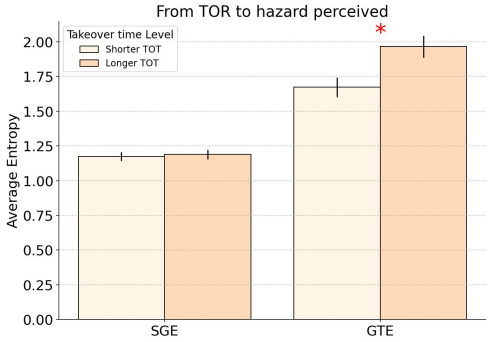
- Takeover time: From the TOR to hazard was perceived (TW2; Figure 5.6c), drivers with shorter TOT exhibited significantly lower GTE. A lower GTE value indicates that drivers' eye movement patterns were more predictable and less uncertain in this TW, showing some degree of consistency with significant lower Entry count in "Center" AOI in this TW.
- Resulting acceleration: No significant differences in either SGE or GTE were found with respect to resulting acceleration.



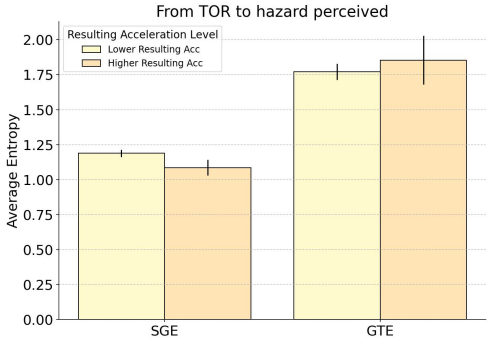
(a) Takeover Time: TW1



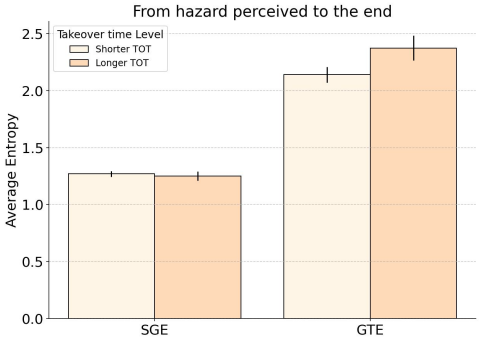
(b) Resulting Acceleration: TW1



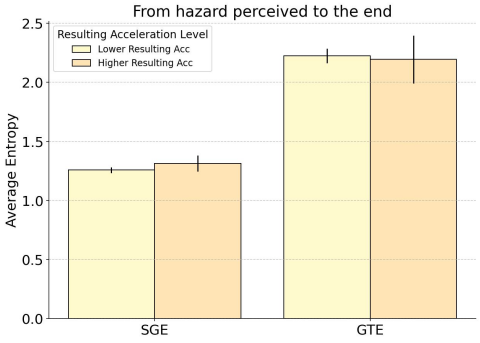
(c) Takeover Time: TW2



(d) Resulting Acceleration: TW2



(e) Takeover Time: TW3



(f) Resulting Acceleration: TW3

Figure 5.6: SGE and GTE of drivers from two Takeover Time groups and two Resulting Acceleration groups across three TWs (error bars represent SEM).

5.4 Discussion and conclusions

In this chapter, differences in eye movement between drivers with different SA levels and takeover performance levels in a semi-autonomous vehicle simulation were investigated. Drivers' eye movement was analyzed and their gaze patterns were represented by four categories of AOI feature. Eye tracking has become a valuable method for examining their cognitive states, providing meaningful insights into their takeover performance.

In the current study, SAGAT score and HPT are two indicators of drivers' SA, with higher SAGAT scores and shorter HPT values representing better SA. The current results do not show any trade-off between reaching a higher SAGAT score and shorter HPT, indicating that a single gaze strategy can support both aspects efficiently. During TW2, both drivers with higher SAGAT score and drivers with shorter HPT showed significantly longer Dwelling time on the "Center" area and shorter time on peripheral areas, suggesting a task-driven likely focus on the location of hazards. Meanwhile, during this TW2, shorter-HPT drivers also exhibited fewer AOI transitions and lower GTE, aligning with a visually predictable and goal-directed eye movement pattern that likely contributed to their faster hazard recognition. After the hazard perceived and till the end of the trial (TW3), shorter-HPT drivers continued the same Dwelling time pattern from TW2, when they had longer Dwelling time on the "Center" AOI and shorter time on "Others" area. During this period, drivers with a higher SAGAT score exhibited more Entry counts into "Center" and "Others" AOIs and significantly higher GTE. These results indicate that they remained active scanning the environment after the hazard has been identified, potentially to maintain a better manual control.

The TOT and Resulting acceleration are two indicators of effective takeover performance, with shorter TOT values and lower resulting acceleration values representing better takeover performance. Similarly, the current results do not show such a trade-off, indicating that a single gaze strategy can support both efficient takeover time and stable vehicle control. Overall, drivers with better takeover performance — whether measured by shorter TOT or lower resulting acceleration — demonstrated fewer gaze entries across multiple AOIs in all three TWs. This pattern suggests that these drivers required fewer visual scans to locate relevant information, resulting in a lower overall number of transitions during the takeover process.

Interestingly, when reviewing all gaze patterns in relation to SA and takeover performance, we observed a inconsistency in gaze behavior during TW2, which is a critical time period immediately following the TOR . Drivers with better SA — represented by both higher SAGAT score and shorter HPT — tended to focus more on the "Center" and less

on the "Other" areas see Figure 5.1c and Figure 5.1d). In contrast, the driver with shorter TOT exhibited the opposite pattern, allocating less attention to the "Center" and more to the "Other" areas (see Figure 5.4c). This inconsistency suggests that SA and quick takeover responses may rely on different gaze strategies during certain time period. While better-SA drivers appear to adopt a more deliberate approach by focusing on central of the view, shorter-TOT drivers may employ a different strategy. Since the moment of the takeover action was not marked as a distinct time point in the current TW segmentation, it is difficult to interpret why shorter-TOT drivers spent less time focusing on the central area. This could possibly be due to shorter-TOT drivers resuming control more quickly and then shifting their attention to peripheral areas to gather additional information. These findings highlight the need to balance gaze strategy depending on whether the goal is to enhance SA or to support fast takeover actions.

Chapter 6

Differences in eye movement between pilots with different Situation Awareness levels and takeover performance levels in an autopilot-assisted instrument landing.

This chapter presents the piloting domain results and discussion of Topic 1: **Differences in eye movement between operators with different Situation Awareness levels and takeover performance levels in a conditionally automated system.**

6.1 Background

The AFCS is one of the vital systems in modern aviation, assisting pilots in maintaining control of the aircraft. The autopilot can provide a better flying quality, improve the primary performance of the aircraft, and lessen pilots' workload at the same time. In some studies, the words "autopilot" and " AFCS" are used interchangeably. Within such automated systems, pilots are required to keep monitoring the system status. In the event of autopilot malfunctions, they must quickly detect the issue and assume manual

control with minimal delay. A lack of understanding of the autopilot system or insufficient awareness of the current flight status can result in serious consequences, as seen in the accidents involving China Airlines Flight 140 and Asiana Airlines Flight 214 (discussed in Chapter 2).

Although automation can significantly decrease pilots' workload, it can also impair their SA level if pilots are either unfamiliar with its functions or overly dependent on it, potentially degrading their manual flying performance. Therefore, it is essential to explore pilots' SA and flight performance during the transition from automated control to manual flying. In the current paper, we use the word "takeover" to describe the transition period where the plane is controlled by an automated system at the beginning and changes to manual flying afterward.

Eye tracking has become a valuable method for examining pilots' cognitive states and behaviors, providing profound insights into their SA and takeover performance. Identifying how eye movement patterns differ between pilots with different SA levels and takeover performance levels is essential for gaining interpretable conclusions that support flight safety. This study investigates four categories of eye-tracking AOI features — Dwelling time, Entry count, SGE and GTE — to explore how pilots with different SA and takeover performance levels visually engage with their environment during the takeover process.

6.2 Methodology

One simulated takeover piloting experiment was conducted to investigate the research question: How does pilots with different SA levels and takeover performance levels move their eyes differently in an autopilot-assisted instrument landing. The details of the piloting experiment is covered in Chapter 3.2.

Pilots' eye movement was collected by eye tracking glasses and their AOI-based eye tracking features were extracted following the method introduced in the Chapter 4.1.3.2. Four categories of AOI features were used to represent pilots' eye movement (Table 4.2): Dwelling time, Entry Count, SGE and GTE. Since each type of feature represents a different aspect of pilots' eye movement, they were discussed one by one in the results section.

To draw insights from the analysis, those eye tracking features were compared between different TWs (defined in Chapter 4.3.2) to examine pilots' eye movement differences between groups as well as changes through time. Consistent with the overall analytical framework introduced in Chapter 4.4.2.1, MANOVA was used to test for multivariate effects, followed by Welch's t-tests for pairwise comparisons when appropriate.

Previous timeline defined three TWs used for the piloting experiment (see Figure 4.9), and the first TW was"Pre-alert". After testing the time length of "60 seconds", "30 seconds" and "10 seconds", it was found that "60 seconds" can best represent the behavior of drivers before the malfunction alert was given. Therefore, the first TW was finally set to be"60 seconds before the alert" in the current piloting experiment.

In all, each piloting experimental trial was separated into three TWs, and all AOI features were extracted separately for each of the three TWs. As described in Chapter 3.2, the landing scenarios included no failure, LOC failure, and autopilot failure. However, not all trials were suitable for analysis. For the investigation of SA, all experimental trials were included. This included no failure trials, as they still allowed examination of how pilots used eye movements to gather situational information across the three TWs, thereby enabling comparisons between different SA groups. In contrast, for the analysis of takeover performance, all no failure trials were excluded, as they did not involve any takeover event by design.

6.3 Results

6.3.1 Pilots' Situation Awareness

Pilots' SA is measured by SAGAT score. For each of the three TWs, separate MANOVA tests using Pillai's Trace were conducted for two SAGAT score groups (the grouping criteria were introduced in Chapter 4.4.1). These analyses examined whether the overall eye movement patterns differed significantly between groups within each TW. When a significant multivariate effect was observed, follow-up Welch's independent t-tests were performed to determine which specific AOI-based features contributed to the observed differences.

Results for higher versus lower SAGAT score groups:

- TW1: Pillai's Trace = 0.119, $F(20, 247) = 1.672, p = 0.038$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between higher and lower SAGAT score groups.
- TW2: Pillai's Trace = 0.161, $F(20, 247) = 1.490, p = 0.091$. No significant multivariate effect was observed, suggesting that eye movement patterns did not differ between higher and lower SAGAT score groups during this time window.

- TW3: Pillai's Trace = 0.130, $F(20, 247) = 1.181, p = 0.277$. No significant multivariate effect was observed, suggesting that eye movement patterns did not differ between higher and lower SAGAT score groups during this time window.

To facilitate interpretation, the follow-up Welch's t-test results are presented in the subsequent sections, organized by feature category (Dwelling Time, Entry Count, and Entropy). Each section reports findings across all three TWs and summarizes the key AOI specific results. Although the multivariate effects in the TW2 and TW3 were not significant, we still report the Welch's t-test results below to keep the consistency.

6.3.1.1 Dwelling time

Pilots with different levels of SA may demonstrate distinct Dwelling time features on various AOIs across different TWs. The dwelling time features contain "Airspeed_DT", "Attitude_DT", "Altitude_DT", "Roll_DT", "Heading_DT", "V Speed_DT", "Power_DT", "Front_DT" and "Others_DT", corresponding to 9 AOIs defined based on experiment scenarios. Each feature represents the percentage of time that pilots spent in the AOI.

Table 6.1 shows the result of Welch's independent t-tests between Dwelling time features of two SAGAT score groups. Figure 6.1 demonstrates the Dwelling time of all AOIs across three TWs, corresponding to Figure 6.1a, Figure 6.1b and Figure 6.1c. Each pair of bars represents the values from two different SAGAT groups of pilots. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. The error bars represent SEM of each bar. Although Welch's t-tests were conducted for SAGAT score during TW2 and TW3 for consistency, their results are not interpreted further because the multivariate effect was not significant.

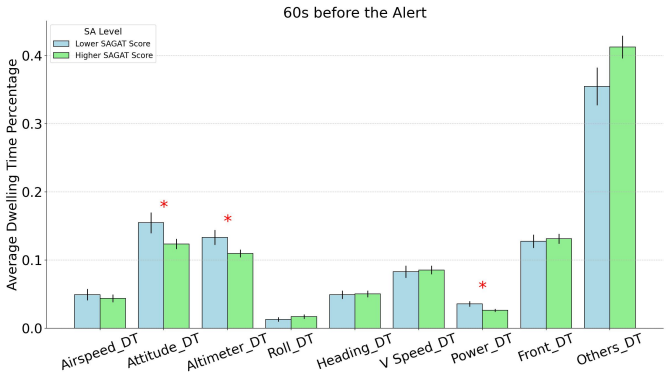
Results In 60s before the alert (TW1; Figure 6.1a), pilots with higher SAGAT score showed significantly less Dwelling time on "Attitude", "altimeter" and "Power" areas compared to pilots with lower SAGAT score.

While some significant differences in Dwelling Time were observed between pilots with higher and lower SAGAT scores, these effects were hard to interpret since the higher SAGAT score group all showed lower Dwelling time on different AOIs. Overall, the observed patterns do not suggest a clear relationship between pilots' SA level and eye movement.

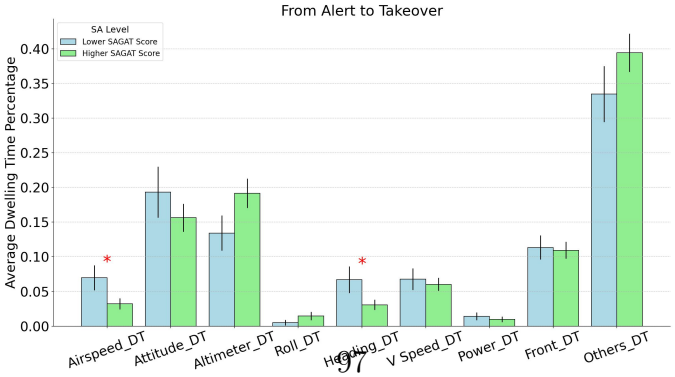
Table 6.1: T-tests for Dwelling time features between pilots from two SAGAT score groups across three TWs (* indicates $p < .05$)

	Airspeed	**Attitude**	**Altimeter**	**Roll**	**Heading**	**V_Speed**	**Power**	**Front**	**Others**
TW1									
T-value	−0.544	−2.064	−2.110	0.836	0.159	0.224	−2.304	0.297	1.829
P-value	0.587	**0.040***	**0.036***	0.404	0.874	0.823	**0.022***	0.767	0.059
Effect size	0.069	0.250	0.257	0.114	0.021	0.029	0.283	0.039	0.237
TW2									
T-value	−2.198	−0.958	1.652	1.081	−2.088	−0.437	−0.679	−0.186	1.239
P-value	**0.029***	0.340	0.100	0.282	**0.038***	0.663	0.498	0.853	0.217
Effect size	0.323	0.146	0.269	0.188	0.302	0.068	0.107	0.029	0.196
TW3									
T-value	−2.397	−1.415	0.904	−0.493	−0.719	1.644	0.568	0.536	−0.053
P-value	**0.018***	0.159	0.368	0.623	0.473	0.102	0.571	0.593	0.958
Effect size	0.356	0.215	0.143	0.075	0.111	0.261	0.089	0.084	0.008

Note: "V_Speed" refers to Vertical Speed. The "DT" in feature names is omitted to save space.



(a) SAGAT Score: TW1



(b) SAGAT Score: TW2



6.3.1.2 Entry count

Pilots with different levels of SA may demonstrate distinct Entry count features on various AOIs across different TWs. The Entry count features contain "Airspeed_EC", "Attitude_EC", "Altitude_EC", "Roll_EC", "Heading_EC", "V Speed_EC", "Power_EC", "Front _EC" and "Others_EC", corresponding to 9 AOIs defined based on experiment scenarios. Each feature represents the percentage of time that pilots spent in the AOI.

Table 6.2 shows the result of Welch's independent t-tests between Entry count features of two SAGAT score groups. Figure 6.2 demonstrates the Dwelling time of all AOIs across three TWs, corresponding to Figure 6.2a, Figure 6.1b and Figure 6.2c. Each pair of bars represents the values from two different SAGAT groups of pilots. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. The error bars represent SEM of each bar. Although Welch's t-tests were conducted for SAGAT score during TW2 and TW3 for consistency, their results are not interpreted further because the multivariate effect was not significant.

Table 6.2: T-tests for Entry count features between pilots from two SAGAT score groups across three TWs (* indicates $p < .05$)

	Airspeed	**Attitude**	**Altimeter**	**Roll**	**Heading**	**V_Speed**	**Power**	**Front**	**Others**
TW1									
T-value	−0.587	−0.032	0.062	0.972	0.509	0.094	−0.876	0.091	1.880
P-value	0.558	0.974	0.951	0.332	0.611	0.926	0.382	0.928	0.061
Effect size	0.075	0.004	0.008	0.132	0.067	0.012	0.112	0.012	0.233
TW2									
T-value	−1.343	−0.809	0.904	0.180	−1.512	−0.648	0.007	−0.181	0.533
P-value	0.181	0.419	0.368	0.858	0.132	0.518	0.994	0.857	0.595
Effect size	0.211	0.125	0.153	0.029	0.219	0.100	0.001	0.029	0.085
TW3									
T-value	−2.366	−0.284	1.277	−0.379	−0.534	1.622	0.608	0.616	0.283
P-value	**0.019***	0.777	0.203	0.705	0.594	0.107	0.543	0.539	0.778
Effect size	0.350	0.044	0.204	0.059	0.082	0.259	0.095	0.095	0.043

Note: "V_Speed" refers to Vertical Speed. The "EC" in feature names is omitted to save space.

Results Across the three TWs, Entry count differences between pilots with higher and lower SAGAT scores were minimal and lacked interpretable patterns. Overall, the En-

try count features were not able to meaningfully differentiate between the eye movement patterns of pilots with higher versus lower SAGAT scores.

6.3.1.3 Entropy

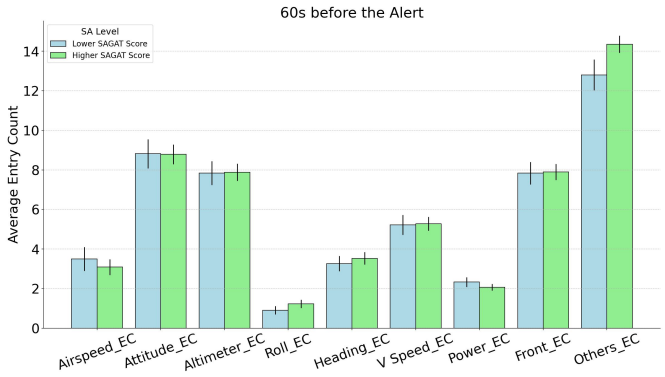
Pilots with different levels of SA may demonstrate distinct SGE and GTE across different TWs.

Table 6.3 shows the result of Welch's independent t-tests for SGE and GTE between two SAGAT score groups. Figure 6.3 demonstrates the Dwelling time of all AOIs across three TWs, corresponding to Figure 6.3a, Figure 6.3b and Figure 6.3c. Each pair of bars represents the values from two different SAGAT groups of pilots. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. The error bars represent SEM of each bar. Although Welch's t-tests were conducted for SAGAT score during TW2 and TW3 for consistency, their results are not interpreted further because the multivariate effect was not significant.

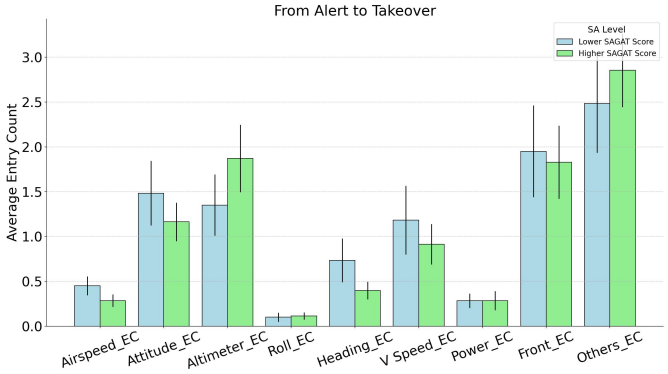
Table 6.3: T-tests for SGE and GTE features between pilots from two SAGAT score groups across three TWs (* indicates $p < .05$)

	SGE	**GTE**
TW1		
T-value	1.434	1.158
P-value	0.152	0.248
Effect size	0.179	0.147
TW2		
T-value	0.639	−0.220
P-value	0.524	0.826
Effect size	0.102	0.035
TW3		
T-value	0.841	0.871
P-value	0.401	0.385
Effect size	0.132	0.135

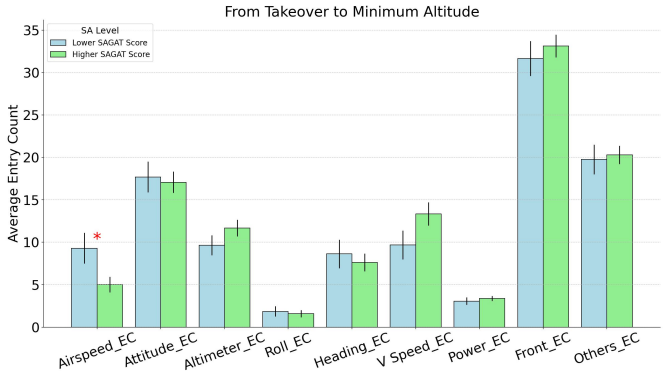
Note: SGE = Stationary Gaze Entropy; GTE = Gaze Transition Entropy.



(a) SAGAT Score: TW1



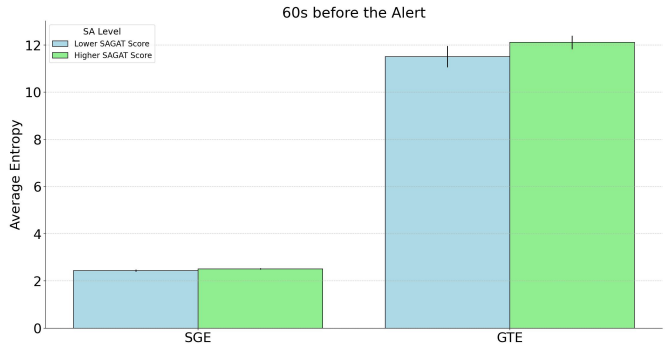
(b) SAGAT Score: TW2



(c) SAGAT Score: TW3

Figure 6.2: Entry count features of pilots from two SAGAT score groups across three TWs (error bars represent SEM).

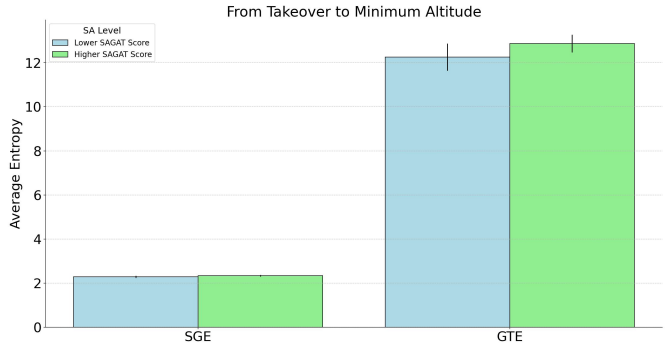
Results The SGE and GTE did not reveal any significant or consistent differences between pilots with higher and lower SAGAT scores across the three TWs. The results suggest that SGE and GTE failed to distinguish significant eye movement patterns of pilots with higher versus lower SAGAT scores.



(a) SAGAT Score: TW1



(b) SAGAT Score: TW2



(c) SAGAT Score: TW3

Figure 6.3: SGE and GTE of pilots from two SAGAT score groups across three TWs (error bars represent SEM).

6.3.2 Pilots' Takeover performance

Pilots' Takeover performance is measured by TOT. For each of the three TWs, separate MANOVA tests using Pillai's Trace were conducted for two TOT groups (the grouping criteria were introduced in Chapter 4.4.1). These analyses examined whether the overall eye movement patterns differed significantly between groups within each TW. When a significant multivariate effect was observed, follow-up Welch's independent t-tests were performed to determine which specific AOI-based features contributed to the observed differences.

Results for shorter versus longer TOT groups:

- TW1: Pillai's Trace = 0.121, $F(20, 158) = 1.812, p = 0.035$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between shorter and longer TOT groups.
- TW2: Pillai's Trace = 0.743, $F(20, 158) = 22.379, p = 0.000$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between shorter and longer TOT groups.
- TW3: Pillai's Trace = 0.196, $F(20, 158) = 1.929, p = 0.137$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between shorter and longer TOT groups.

To facilitate interpretation, the follow-up Welch's t-test results are presented in the subsequent sections, organized by feature category (Dwelling Time, Entry Count, and Entropy). Each section reports findings across all three TWs and summarizes the key AOI specific results.

6.3.2.1 Dwelling time

Pilots with different levels of TOT may demonstrate distinct Dwelling time features on various AOIs across different TWs. The dwelling time features contain "Airspeed_DT", "Attitude_DT", "Altitude_DT", "Roll_DT", "Heading_DT", "V Speed_DT", "Power_DT", "Front_DT" and "Others_DT", corresponding to 9 AOIs defined based on experiment scenarios. Each feature represents the percentage of time that pilots spent in the AOI.

Table 6.4 shows the results of Welch's independent t-tests between Dwelling time features of two TOT groups. Figure 6.4 demonstrates the Dwelling time of all AOIs across

three TWs, corresponding to Figure 6.4a, Figure 6.4b and Figure 6.4c. Each pair of bars represents the values from two different TOT groups of pilots. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. The error bars represent SEM of each bar.

Table 6.4: T-tests for Dwelling time features between pilots from two Takeover time groups across three TWs (* indicates $p < .05$)

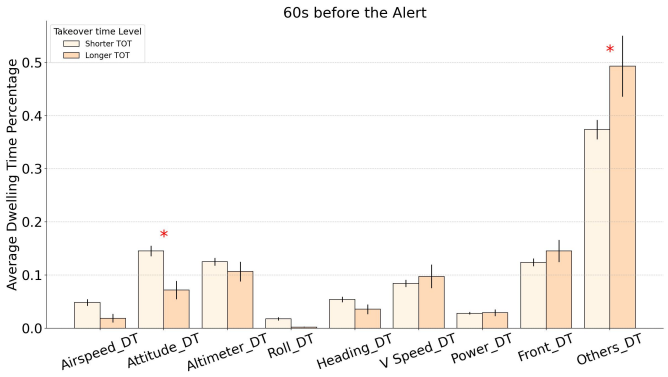
	Airspeed	**Attitude**	**Altimeter**	**Roll**	**Heading**	**V_Speed**	**Power**	**Front**	**Others**
TW1									
T-value	1.776	2.659	0.915	1.963	1.241	−0.676	−0.149	−1.040	−2.243
P-value	0.078	**0.009***	0.362	0.051	0.216	0.500	0.882	0.300	**0.026***
Effect size	0.490	0.694	0.211	0.590	0.325	0.140	0.033	0.228	0.479
TW2									
T-value	1.830	1.462	0.886	0.580	−0.118	−2.292	−1.543	−3.139	−0.154
P-value	0.069	0.146	0.378	0.563	0.907	**0.023***	0.125	**0.002***	0.878
Effect size	0.549	0.412	0.252	0.173	0.032	0.555	0.435	0.772	0.039
TW3									
T-value	1.447	1.122	0.140	0.834	1.294	1.383	−1.137	−1.249	−1.548
P-value	0.150	0.263	0.889	0.405	0.197	0.169	0.258	0.213	0.123
Effect size	0.377	0.245	0.033	0.232	0.335	0.378	0.221	0.298	0.324

Note: "V_Speed" refers to Vertical Speed. The "DT" in features names is omitted to save space.

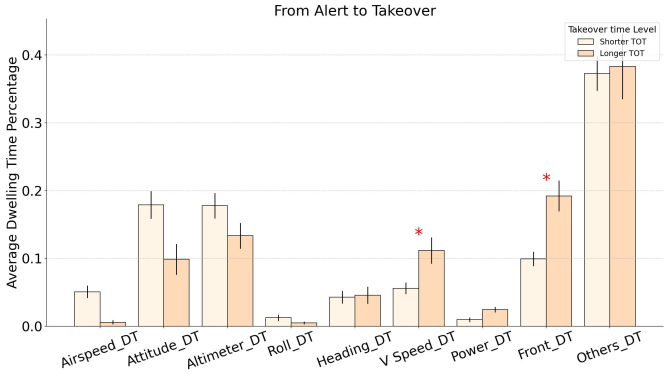
Results During the first TW (60s before the alert; Figure 6.4a), pilots with shorter TOT had significantly more Dwelling time in the "Attitude" area (see Figure 4.5, where the visual cue of the malfunction alert appears), and significantly less dwelling time in the "Others" area. This pattern suggests that faster-responding pilots were already monitoring the area that would soon present the alert cue and, by extension, were less distracted by peripheral or non-critical information.

After the alert and before the takeover (TW2; Figure 6.4b), pilots with shorter TOT spent significantly less time in two of the AOIs: "Vertical Speed" and "Front". One interpretation is that these pilots reallocated their visual attention to other instruments (e.g., attitude, heading) to confirm aircraft state and plan their control inputs, rather than lingering on vertical-speed trends or the forward scene.

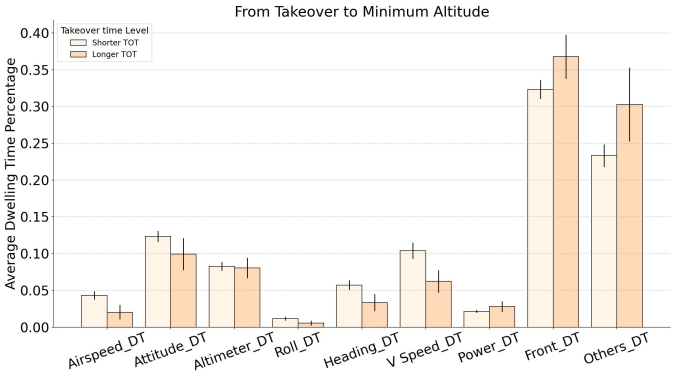
No significant results were found in the last TW (TW3; Figure 6.4c), meaning pilots from longer-TOT group and shorter-TOT group showed similar Dwelling time on all AOIs.



(a) Takeover Time: TW1



(b) Takeover Time: TW2



(c) Takeover Time: TW3

Figure 6.4: Dwelling time features of pilots from two Takeover Time groups groups across three TWs (error bars represent SEM).

6.3.2.2 Entry count

Pilots with different levels of TOT may demonstrate distinct Entry count features on various AOIs across different TWs. The Entry count features contain "Airspeed_EC", "Attitude_EC", "Altitude_EC", "Roll_EC", "Heading_EC", "V Speed_EC", "Power_EC", "Front_EC" and "Others_EC", corresponding to 9 AOIs defined based on experiment scenarios. Each feature represents the percentage of time that pilots spent in the AOI.

Table 6.5 shows the result of Welch's independent t-tests between Entry count features of two TOT groups. Figure 6.5 demonstrates the Dwelling time of all AOIs across three TWs, corresponding to Figure 6.5a, Figure 6.5b and Figure 6.5c. Each pair of bars represents the values from two different TOT groups of pilots. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. The error bars represent SEM of each bar.

Table 6.5: T-tests for Entry Count features between pilots from two Takeover time groups across three TWs (* indicates $p < .05$)

	Airspeed	**Attitude**	**Altimeter**	**Roll**	**Heading**	**V_Speed**	**Power**	**Front**	**Others**
TW1									
T-value	1.699	3.177	1.340	1.993	0.950	−0.213	0.225	−0.727	0.629
P-value	0.091	**0.002***	0.182	**0.048***	0.343	0.832	0.823	0.469	0.531
Effect size	0.463	0.829	0.332	0.596	0.239	0.046	0.053	0.165	0.147
TW2									
T-value	−0.439	−8.816	−8.810	−3.124	−8.111	−10.043	−8.592	−13.327	−12.775
P-value	0.661	**0.000***	**0.000***	**0.002***	**0.000***	**0.000***	**0.0001***	**0.000***	**0.000***
Effect size	0.090	1.066	1.080	0.514	1.041	1.248	1.009	1.569	1.601
TW3									
T-value	1.823	3.120	1.951	1.038	1.712	2.113	0.614	2.880	1.982
P-value	0.070	**0.002***	0.053	0.301	0.089	**0.036***	0.540	**0.004***	**0.049***
Effect size	0.504	0.829	0.533	0.292	0.456	0.571	0.129	0.734	0.509

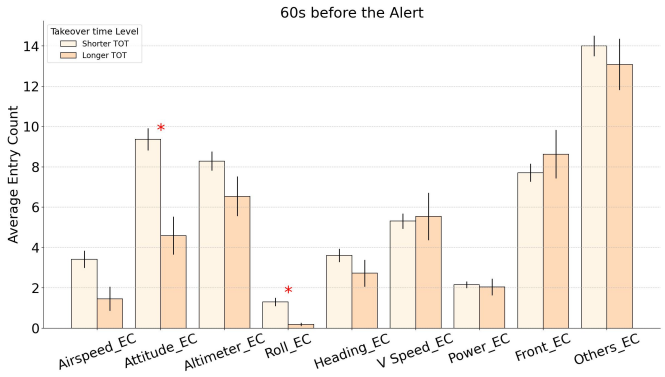
Note: "V_Speed" refers to Vertical Speed. The "EC" in feature names was omitted to save space.

Results In the 60s before the alert (TW1, Figure 6.5a), pilots with shorter TOT had significantly more Entry count in the "Attitude" area, where the visual cue of the malfunction alert appears. This aligns with the earlier Dwelling Time results, further indicating that faster-responding pilots had already been monitoring the alert-relevant region beforehand.

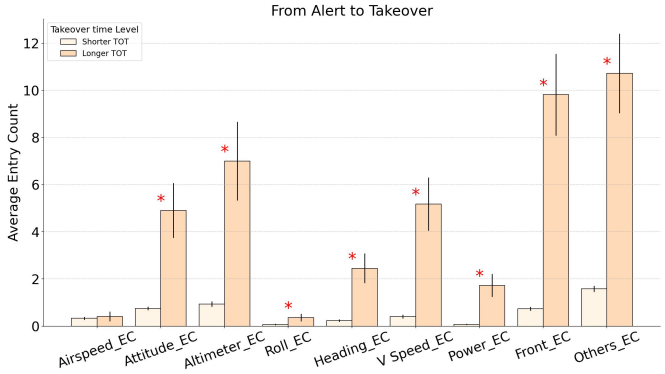
After the alert was issued and prior to the takeover action (TW2; Figure 6.5b), pilots

with shorter TOT exhibited significantly fewer Entry counts than those with longer TOT in all AOIs except "Airspeed". The widespread significance in most AOIs makes this pattern particularly notable. This suggests that shorter-TOT pilots conducted less visual scanning after detecting the alert—possibly because they were more familiar with the system and could quickly identify relevant information without extensive scanning.

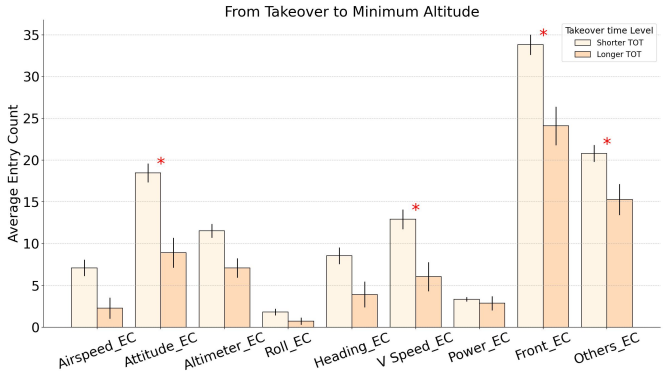
From the moment of takeover to the minimal altitude (TW3; Figure 6.5c), pilots with shorter TOT showed significantly more Entry count in four AOIs: "Attitude", "Vertical speed", "Front" and "Others". These areas include both information critical areas and also the "Others", which means the shorter-TOT pilots kept scanning the whole view during this TW. These areas include both critical flight information and peripheral views, suggesting that shorter-TOT group pilots engaged in more active visual scanning during this TW. This broader search behavior may reflect ongoing monitoring and adjustment as they manually control the aircraft after the takeover.



(a) Takeover Time: TW1



(b) Takeover Time: TW2



(c) Takeover Time: TW3

Figure 6.5: Entry count features of pilots from two Takeover Time groups groups across three TWs (error bars represent SEM).

6.3.2.3 Entropy

Pilots with different levels of TOT may demonstrate distinct SGE and GTE across different TWs.

Table 6.6 shows the result of Welch's independent t-tests for SGE and GTE between two TOT groups. Figure 6.6 demonstrates the Dwelling time of all AOIs across three TWs, corresponding to Figure 6.6a, Figure 6.6b and Figure 6.6c. Each pair of bars represents the values from two different TOT groups of pilots. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. The error bars represent SEM of each bar.

Table 6.6: T-tests for SGE and GTE features between pilots from two Takeover Time groups across three TWs (* indicates $p < .05$)

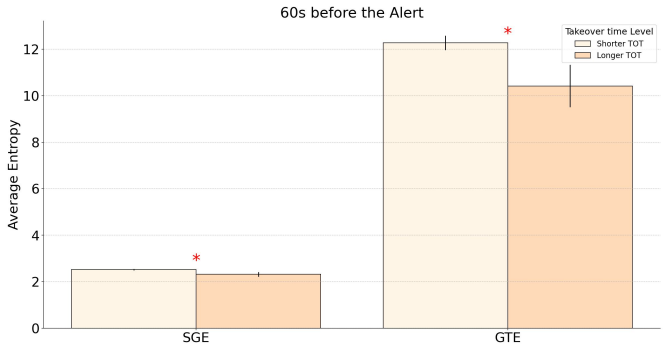
	SGE	**GTE**
TW1		
T-value	2.321	2.067
P-value	**0.021***	**0.040***
Effect size	0.485	0.454
TW2		
T-value	−7.217	−17.32
P-value	**0.000***	**0.000***
Effect size	2.016	2.958
TW3		
T-value	0.801	2.529
P-value	0.423	**0.012***
Effect size	0.197	0.582

Note: SGE = Stationary Gaze Entropy; GTE = Gaze Transition Entropy.

Results In the 60 seconds before the alert (TW1, Figure 6.6a), shorter-TOT pilots had significantly higher SGE and GTE than longer-TOT pilots. This indicates that shorter-TOT pilots engaged in more dispersed and complex visual scanning and monitoring. These findings show some degree of consistency with their significantly higher Entry counts in two AOIs observed during this TW, as discussed in the previous section. The significant differences in both entropy metrics indicated that shorter-TOT pilots distributed their gaze more diversely and shifted attention more frequently across AOIs than longer-TOT pilots.

After the alert was issued and prior to the takeover action (TW2, Figure 6.6b), the pattern reversed: shorter-TOT pilots demonstrated significantly lower SGE and GTE than longer-TOT pilots. This suggests that shorter-TOT pilots engaged in more focused and predictable searching during this period. Their familiarity with the procedure for disengaging the autopilot likely enabled them to quickly locate the relevant controls, leading to more focused gaze behavior and lower entropy values.

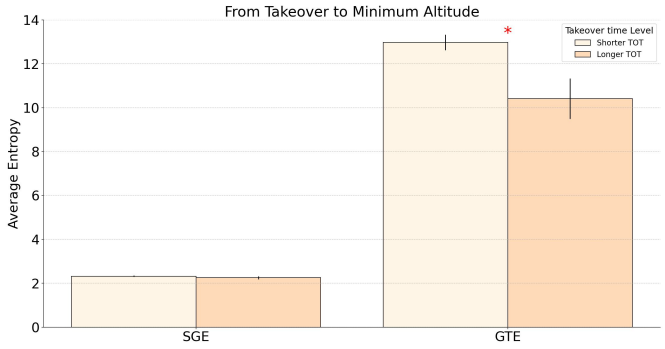
From the moment of takeover to the minimal altitude (TW3; Figure 6.6c), the trend shifted back to resemble TW1. Shorter-TOT pilots showed significantly higher GTE than longer-TOT pilots, indicating that shorter-TOT pilots engaged in more complex visual monitoring and adjustment as they were manually control the aircraft after the takeover. This shows some degree of consistency with the Entry count results of TW3 as well, where shorter-TOT pilots had significantly more entries into four different AOIs, further supporting the interpretation of heightened visual monitoring during manual flying in TW3.



(a) SAGAT Score: TW1



(b) SAGAT Score: TW2



(c) SAGAT Score: TW3

Figure 6.6: SGE and GTE of pilots from two Takeover time groups across three TWs (error bars represent SEM).

6.4 Discussion and conclusions

In this chapter, differences in eye movement between pilots with different Situation Awareness levels and takeover performance levels in an autopilot-assisted instrument landing were explored. Pilots' eye movement was analyzed and their gaze patterns were represented by four categories of AOI feature. Eye tracking has become a valuable method for examining pilots' eye movement patterns, providing profound insights into their takeover performance.

Overall, the results across all four eye movement metrics—Dwelling time, Entry count, SGE and GTE—consistently revealed meaningful differences in eye movement behavior between pilots with shorter and longer TOT. Pilots with shorter TOT demonstrated more focused and task-relevant eye movement behavior prior to the alert (TW1), as evidenced by increased Dwelling time and Entry count in critical AOIs (e.g. "Attitude" which displays the malfunction alerts), as well as higher entropy values that suggest dynamic and dispersed scanning. Subsequently, immediately after the alert (TW2), shorter-TOT pilots exhibited a more focused and predictable visual pattern with less visual scanning among AOIs, as evidenced by significantly less Entry count in most AOIs, as well as lower SGE and GTE. Those characteristics indicated rapid identification of the necessary controls with minimal visual searching. Following takeover (TW3), their eye movement patterns once again became more dynamic and complex, with increased GTE and frequent engagement with multiple AOIs—supporting the interpretation that shorter-TOT pilots maintained stronger visual awareness and control during manual flying.

However, as for the SA, no consistent or interpretable patterns were found that reliably distinguished pilots with higher versus lower SAGAT scores. While a few significant differences in Dwelling time emerged in specific AOIs during certain TWs, these findings were isolated and lacked a coherent trend across different periods of time. Collectively, these results suggest that traditional AOI features may have limited sensitivity in capturing differences in pilots SA. Additionally, the lack of clear findings may also be attributed to the design of the SAGAT questions. On one hand, it was challenging to develop questions that were both comprehensive and reasonable to the landing scenarios used in this study. This difficulty has been rarely addressed in previous research. On the other hand, all questions about the entire landing process were presented only at the end of the trial to avoid disrupting the experiment. This delay in questioning may have introduced memory-related limitations, potentially reducing the accuracy of pilots' responses and weakening the measured relationship between eye movements and pilots' SA.

Together, these findings offer a deeper understanding of how SA and TOT are as-

sociated with pilots' eye movement behaviors in different TWs, and also highlight the complementary nature of the four AOI features in capturing distinct aspects of pilots' eye movements.

Chapter 7

Differences in eye movement between drivers from different age groups during the takeover process in a semi-autonomous driving system

This chapter presents the driving domain results and discussion of Topic 2: **Differences in eye movement between operators from different population groups during the takeover process in a conditionally automated system.**

7.1 Background

In level 3 autonomous vehicles, the transfer process from autonomous control to manual control — referred to as the takeover process — is challenging, particularly for older drivers. Many research studies have demonstrated that older people exhibit decreased takeover performance compared to other age groups, which is usually assessed by metrics such as takeover time and takeover quality. As individuals get older, declines in cognitive function, vision, and decision-making abilities may occur. Especially, changes in their visual perception can also cause trouble when they face a sudden takeover request.

Given the increasing presence of older drivers and semi-autonomous vehicles on today's roads, substantial research efforts have been devoted to understanding their behavior and challenges. Eye movement behavior can provide valuable conclusions into how drivers

gather and process visual information during such transitions. By examining age-related differences in eye movement patterns, we can gain insight into how drivers from different age groups monitor the driving conditions and respond to takeover prompts. These insights could improve the development of more customized and effective TOR designs in the future.

Additionally, current studies usually treat older drivers as one homogeneous group, and typically recruit younger-old drivers in their experiment, who possess different characteristics compared to older old people, thereby overlooking potential differences among various subgroups within this demographic. There are various ways of categorizing older people [55] and the current study used the definition of three like-stage subgroups: the young-old (65–74 years old), the middle-old (75–84 years old), and the old-old (over 85 years old) [166]. To better understand the effects of aging on visual behavior, this study specifically recruited drivers aged 75 and above, including the middle-old and old-old subgroups, as discussed in Chapter 3.1.3.

7.2 Methodology

One simulated takeover driving experiment was conducted to investigate the research question: How does drivers from different age groups move their eyes differently in a semi-autonomous driving system. The details of the driving experiment is covered in Chapter 3.1. As discussed in Chapter 3.1.3, there were three age groups recruited for the current driving experiment, younger, middle-aged and older drivers. After processing the data, it was found that the eye movement of the younger and middle-aged groups shared high similarity. Also, the middle-aged group also exhibited inconsistent characteristics, at times resembling the younger group and at other times the older group. Therefore, to make the results easier to interpret, the younger and middle-aged groups were combined into one single age group in the results section.

Drivers' eye movement was collected by eye tracking glasses and their AOI-based eye tracking features were extracted following the steps introduced in Chapter 4.1.3.1. Four categories of AOI features were used to represent drivers' eye movement (Table 4.2): Dwelling time, Entry Count, SGE and GTE. Since each type of feature represents a different aspect of drivers' eye movement, they were discussed one by one in the results section. After processing the data, we found that the dwelling time and entry counts of three mirrors: Rear-mirror, Left mirror, and Right mirror were relatively low. Therefore, these three AOIs were combined into one single AOI labeled as the "Mirrors".

To draw insights from the analysis, eye tracking features were compared between different TWs (defined in Chapter 4.3.1) to examine drivers' eye movement differences between

age groups as well as changes through time. Consistent with the overall analytical framework introduced in Chapter 4.4.2.1, MANOVA was used to test for multivariate effects, followed by Welch's t-tests for pairwise comparisons when appropriate.

In the data analysis process, some revisions were made towards TW separation. Previous timeline defined three TWs used for the driving experiment (see Figure 4.8), and the first TW was "Pre-TOR". After testing time periods of "15 seconds", "10 seconds" and "5 seconds", it was found that "10 seconds" can best represent the behavior of drivers before the TOR. Therefore, this TW was finally set to be "10 seconds before the TOR" in the current driving experiment.

In all, each driving experimental trial was separated into three TWs, and all AOI features were extracted separately for each of the three TWs. As described in Chapter 3.1, the driving scenarios included manual, automation without NDRT and automation with NDRT, where the NDRT included two visual tasks and one auditory task. However, not all trials were suitable for analysis. For example, since the AOI setups in the driving experiment assumed drivers were viewing the front monitor, trials involving visual distractions were excluded. Because these trials required drivers to engage with a separate side laptop, which meant they were not looking at the front screen where the AOIs were defined. In the end, for the investigation of drivers' age, all remaining experimental trials were included. This included manual trials, as they still allowed examination of how drivers used eye movements to gather situational information across the three TWs, thereby enabling comparisons between different age groups.

7.3 Results

All drivers recruited for the current study are separated into to age groups: younger and older. For each of the three TWs, separate MANOVA tests using Pillai's Trace were conducted for two age groups. These analyses examined whether the overall eye movement patterns differed significantly between age groups within each TW. When a significant multivariate effect was observed, follow-up Welch's independent t-tests were performed to determine which specific AOI-based features contributed to the observed differences.

Results for younger versus older groups:

- TW1: Pillai's Trace $= 0.121$, $F(11, 467) = 5.820, p = 0.000$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between younger and older groups.

- TW2: Pillai's Trace = 0.154, $F(11, 467) = 7.730, p = 0.000$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between younger and older groups.
- TW3: Pillai's Trace = 0.193, $F(11, 467) = 10.130, p = 0.000$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between younger and older groups.

To facilitate interpretation, the follow-up Welch's t-test results are presented in the subsequent sections, organized by feature category (Dwelling Time, Entry Count, and Entropy). Each section reports findings across all three TWs and summarizes the key AOI specific results.

7.3.1 Dwelling time

Drivers with different ages might exhibit different Dwelling time features on various AOIs across different TWs. The dwelling time features contain "Mirrors_DT", "Center_DT", "Dashboard_DT" and "Others_DT", corresponding to the 6 AOIs defined based on experiment scenarios, with three mirrors combined into a single AOI named "Mirrors". Each feature represents the percentage of time that drivers spent in the AOI.

Table 7.1 shows the results of Welch's independent t-tests between Dwelling time features of two age groups. Figure 7.1 demonstrates the Dwelling time of all AOIs across three TWs, corresponding to Figure 7.1a, Figure 7.1b and Figure 7.1c. Each pair of bars represents the values from two different groups of drivers. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this Dwelling time feature between the two age groups. The error bars represent SEM of each bar.

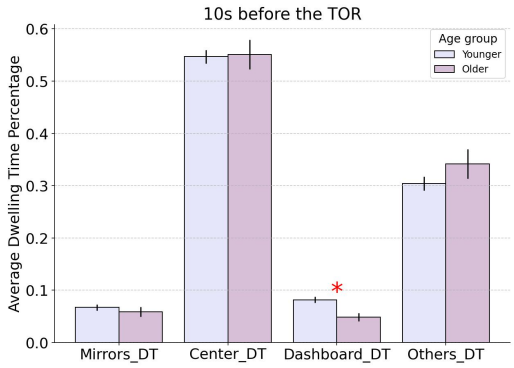
Results From 10 seconds before the TOR till hazard perceived (TW1 and TW2; Figure 7.1a and Figure 7.1b), older drivers had significantly less dwelling time on "Dashboard" compared to younger drivers. And from hazard perception to the end (TW3; Figure 7.1c), Older drivers had significantly less dwelling time on "Dashboard" and "Mirrors" compared to younger drivers.

This result implied that older drivers paid less attention to certain information critical areas, potentially failing to adequately monitor key vehicle status throughout the process.

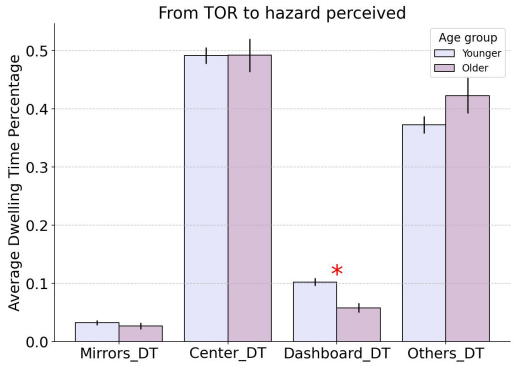
Table 7.1: T-tests for Dwelling times features between drivers from two age groups across three TWs (* indicates $p < .05$)

	Younger vs. Older Drivers			
	Mirrors_DT	**Center_DT**	**Dashboard_DT**	**Others_DT**
TW1				
T-value	0.652	−0.144	2.657	−1.225
P-value	0.515	0.886	**0.008***	0.221
Effect size	0.081	0.016	0.342	0.140
TW2				
T-value	0.664	−0.003	3.059	−1.491
P-value	0.507	0.998	**0.002***	0.137
Effect size	0.086	0.000	0.404	0.172
TW3				
T-value	2.935	−0.472	3.529	−1.678
P-value	**0.003***	0.637	**0.000***	0.094
Effect size	0.387	0.055	0.453	0.199

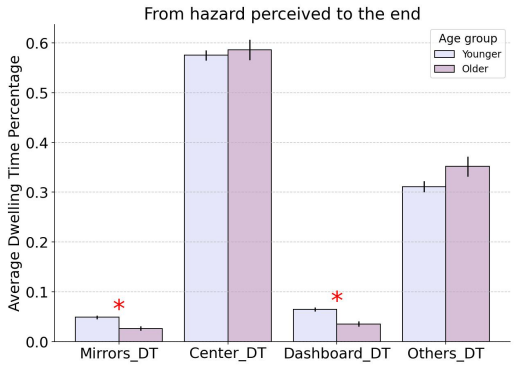
This reduced attention may be attributed to age-related cognitive decline or decreased familiarity with the simulated driving environment.



(a) TW1



(b) TW2



(c) TW3

Figure 7.1: Dwelling time features of drivers from two age groups across three TWs (error bars represent SEM).

7.3.2 Entry count

Drivers with different ages might exhibit different Entry count features on various AOIs across different TWs. The Entry count features contain "Mirrors_EC", "Center_EC", "Dashboard_EC", "Screen_EC" and "Others_EC", corresponding to the 6 AOIs defined based on experiment scenarios, with three mirrors combined into a single AOI named "Mirrors". Each feature represents the percentage of time that drivers spent in the AOI.

Table 7.2 shows the results of Welch's independent t-tests between Entry count features of two age groups. Figure 7.2 demonstrates the Entry count of all AOIs across three TWs, corresponding to Figure 7.2a, Figure 7.2b and Figure 7.2c. Each pair of bars represents the values from two different groups of drivers. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this Entry count feature between the two age groups. The error bars represent SEM of each bar.

Table 7.2: T-tests for Entry count features between drivers from two age groups across three TWs (* indicates $p < .05$)

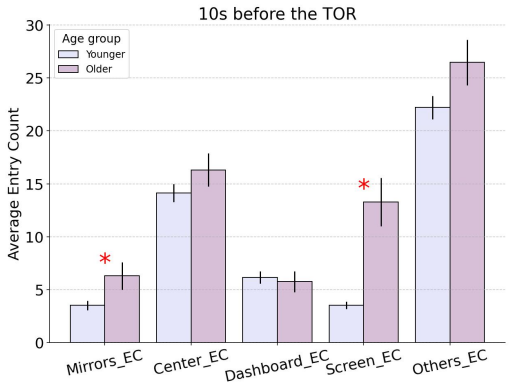
	Younger vs. Older Drivers				
	Mirrors_EC	**Center_EC**	**Dashboard_EC**	**Screen_EC**	**Others_EC**
TW1					
T-value	−2.572	−1.143	0.310	−7.371	−1.707
P-value	**0.010***	0.254	0.757	**0.000***	0.089
Effect size	0.260	0.137	0.038	0.597	0.200
TW2					
T-value	−3.657	−3.312	−1.426	−8.114	−4.086
P-value	**0.000***	**0.001***	0.154	**0.000***	**0.000***
Effect size	0.302	0.377	0.165	0.607	0.439
TW3					
T-value	−1.825	−2.271	−0.658	−8.063	−2.457
P-value	0.069	**0.024***	0.511	**0.000***	**0.014***
Effect size	0.179	0.259	0.077	0.615	0.278

Results From 10 seconds before the TOR (TW1; Figure 7.2a), older drivers had significantly more entries into the AOIs "Mirrors" and "Screen". The higher Entry count for the

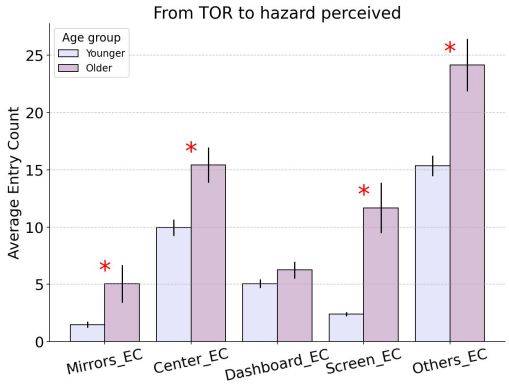
"Screen" area suggests that their gaze shifted in and out of this region more frequently, which may indicate greater visual distraction or divided attention during the autonomous driving phase.

From TOR till hazard perceived (TW2; Figure 7.2b), older drivers had significantly more Entry counts in all areas except the "Dashboard", suggesting a more active visual search behavior immediately following the TOR. This could reflect uncertainty in determining their next action or difficulty in identifying the hazard quickly.

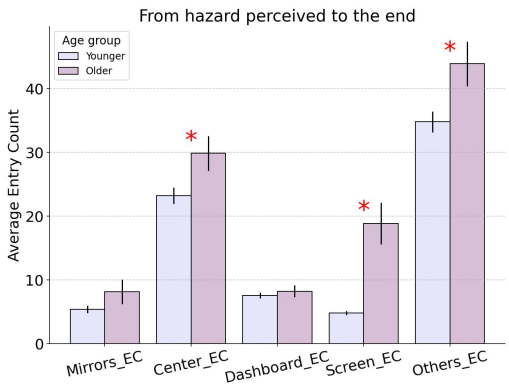
And from hazard perception to the end (TW3; Figure 7.2c), older drivers still had more Entry count in "Center", "Screen" and "Others" areas, showing that they were still actively scanning different areas.



(a) TW1



(b) TW2



(c) TW3

Figure 7.2: Entry count features of drivers from two age groups across three TWs (error bars represent SEM).

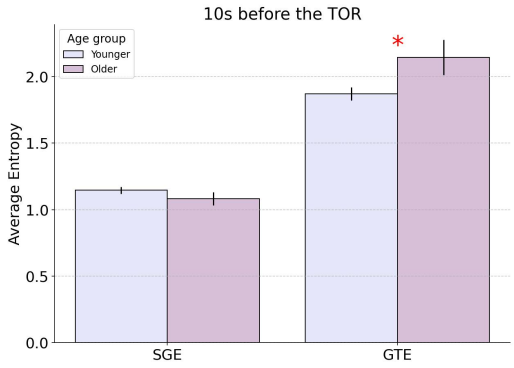
7.3.3 Entropy

Drivers with different ages might exhibit different SGE and GTE across different TWs.

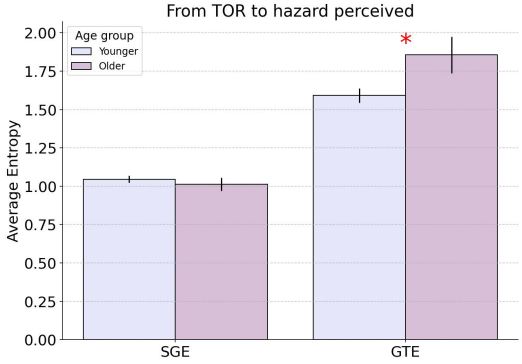
Table 7.3 shows the results of Welch's independent t-tests between SGE and GTE of two age groups. Figure 7.3 demonstrates the SGE and GTE across three TWs, corresponding to Figure 7.3a, Figure 7.3b and Figure 7.3c. Each pair of bars represents the values from two different groups of drivers. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in SGE and GTE between the two age groups. The error bars represent SEM of each bar.

Table 7.3: T-tests for SGE and GTE between drivers from two age groups across three TWs (* indicates $p < .05$)

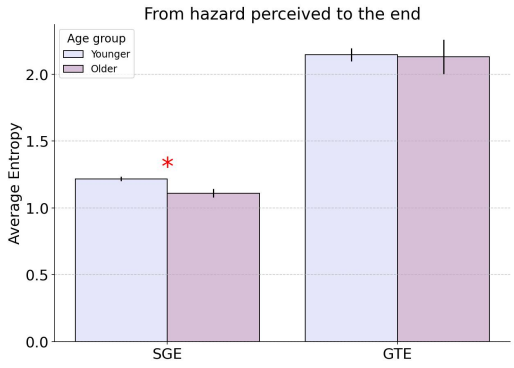
	SGE	**GTE**
TW1		
T-value	1.111	−2.298
P-value	0.267	**0.022***
Effect size	0.130	0.242
TW2		
T-value	0.649	−2.384
P-value	0.517	**0.018***
Effect size	0.077	0.256
TW3		
T-value	2.720	0.121
P-value	**0.007***	0.904
Effect size	0.326	0.013



(a) TW1



(b) TW2



(c) TW3

Figure 7.3: SGE and GTE of drivers from two age groups across three TWs (error bars represent SEM).

Results From 10 seconds before the TOR till hazard perceived (TW1 and TW2; Figure 7.3a and Figure 7.3b), older drivers had significantly higher GTE values than younger drivers. This indicates more complex transitions among different AOIs, suggesting a more uncertain scanning pattern. This result shows some degree of consistency with the result of Entry count during TW1 when older drivers did enter multiple AOIs significantly more frequently than younger drivers. And from hazard perception to the end (TW3; Figure 7.3c), older drivers had a significantly low SGE, indicating a more concentrated gaze distribution during this time period. This shift may reflect a narrowed attention strategy or increased cognitive effort to process relevant information once control had been resumed.

7.4 Discussion and conclusions

This study examined age-related differences in drivers' eye movement behavior during the takeover process in conditionally automated vehicles, focusing on four categories of eye-tracking features: Dwelling Time, Entry count, SGE, and GTE. The results consistently showed that older drivers exhibited distinct visual scanning patterns compared to younger drivers across different TWs of the takeover process.

In terms of Dwelling time, the results showed that older drivers exhibited reduced visual attention to information critical areas, proved as they spent significantly less time fixating on areas like "Dashboard" and "Mirrors" . This situation reflected their diminished information gathering capabilities potentially influenced by age-related cognitive decline or reduced familiarity with the simulated environment.

The Entry count analysis further revealed that older drivers made significantly more frequent gaze entries into multiple areas. From TW1 to TW3, older drivers always demonstrated significantly more entries in multiple AOIs compared to younger drivers, rather than less. Especially in TW2, they continued to exhibit significantly higher Entry counts across almost all AOIs, indicative of a more intensive visual search following the TOR. These results revealed their active scanning activities, which may point to compensatory behaviors to overcome reduced processing efficiency or to confirm their driving decisions.

The Entropy values generally supported these findings. Older drivers showed significantly higher GTE in TW1 and TW2, indicating more complex transitions between AOIs. and reflecting unpredictable but possibly inefficient search strategy. In TW3, their SGE was significantly lower, suggesting a more focused gaze distribution during the manual driving phase.

Together, these findings suggest that older drivers engage in more extensive visual scanning prior to and during the takeover transition but may ignore critical areas. Their behavior points to potential challenges in prioritizing information sources effectively under time pressure. As reviewed in Chapter 2.3.2, previous research has also demonstrated that older drivers typically show more frequent fixations and a reduced capacity for efficient information gathering. This pattern is consistent with the higher Entry counts and GTE observed in the current study. Moreover, earlier findings have indicated that older drivers adopt more centrally focused gaze strategies during both conventional driving and takeover scenarios, which is further supported by the lower SGE results in the present analysis.

In all, these age-related differences reveal the importance of adaptive interface design and tailored takeover strategies that accommodate older drivers' own visual and cognitive processing characteristics. Supporting older drivers with clearer visual cues, simplified displays, or pre-alert information may improve their SA and takeover performance in future automated driving systems.

Chapter 8

Differences in eye movement between pilots from different flight experience groups during the takeover process in an autopilot-assisted instrument landing.

This chapter presents the piloting domain results and discussion of Topic 2: **Differences in eye movement between operators from different population groups during the takeover process in a conditionally automated system.**

8.1 Background

Modern aircraft heavily rely on the AFCS, which plays a critical role in supporting pilots with aircraft control tasks. It can enhance flight stability, improve flying performance, and decrease pilots' workload; the autopilot function contributes significantly to safer and more efficient operations. Despite these benefits, pilots are still responsible for continuously monitoring the system. When automation failures occur, they need to do a manual takeover with a minimum time delay.

The takeover process from autopilot to manual flying could be challenging. To investigate how pilots with different experience move their eyes differently during the takeover

process of autopilot, an piloting experiment was conducted. In the current study, two groups of pilots were recruited. As discussed in Chapter 3.2.3, the expert group is composed of pilots with their Instrument rating certificate, enabling them to operate certain autopilot functions independently, and the learner group is composed of pilots without the Instrument rating but with sufficient knowledge of ILS landing to complete the experimental tasks under the guidance of the experimenter.

Investigating eye movement differences between pilots from different experience levels allows us to identify how they visually engage with the cockpit environment, which is critical to maintaining safety and performance during demanding phases of flight. Eye tracking techniques have been widely used in aviation studies to examine pilots' cognitive processes and performance. In the current study, pilots' eye movement behavior reflects the pilot's SA and attention allocation. Experienced pilots are expected to develop more efficient scanning patterns and information-gathering strategies.

By comparing eye-tracking AOI features — including Dwelling time, Entry count, SGE and GTE — between learner and expert groups across different time periods in the takeover process, we can gain valuable insights into how experience shapes visual attention and ultimately impacts takeover performance. The **Results** sections under each category of AOI features state the key findings. Ultimately, the insights gained from this study are expected to inform and enhance pilot training programs for managing manual takeovers.

8.2 Methodology

One simulated takeover piloting experiment was conducted to investigate the research question: How does pilots from different experience group move their eyes differently in an autopilot-assisted instrument landing. The details of the piloting experiment is covered in Chapter 3.2. As discussed in Chapter 3.2.3, there were two experience groups recruited for the current piloting experiment.

Pilots' eye movement was collected by eye tracking glasses and their AOI-based eye tracking features were extracted following the steps introduced in the Chapter 4.1.3.2. Four categories of AOI features were used to represent pilots' eye movement (Table 4.2): Dwelling time, Entry Count, SGE and GTE. Since each type of feature represents a different aspect of pilots' eye movement, they were discussed one by one in the results section.

To draw insights from the analysis, those eye tracking features were compared between different TWs (defined in Chapter 4.3.2) to examine pilots' eye movement differences between experience groups as well as changes through time. Consistent with the overall

analytical framework introduced in Chapter 4.4.2.1, MANOVA was used to test for multivariate effects, followed by Welch's t-tests for pairwise comparisons when appropriate.

Previous timeline defined three TWs used for the piloting experiment (see Figure 4.9), and the first TW was"Pre-alert". After testing the time length of "60 seconds", "30 seconds" and "10 seconds", it was found that "60 seconds" can best represent the behavior of drivers before the malfunction alert was given. Therefore, the first TW was finally set to be"60 seconds before the alert" in the current piloting experiment.

In all, each piloting experimental trial was separated into three TWs, and all AOI features were extracted separately for each of the three TWs. As described in Chapter 3.2, the landing scenarios included no failure, LOC failure, and autopilot failure. However, not all trials were suitable for analysis. For the investigation of experience, all experimental trials were included. This included no failure trials, as they still allowed examination of how pilots used eye movements to gather situational information across the three TWs, thereby enabling comparisons between different experience groups.

8.3 Results

All pilots recruited for the current study are separated into to experience groups: learner and expert. For each of the three TWs, separate MANOVA tests using Pillai's Trace were conducted for two experience groups. These analyses examined whether the overall eye movement patterns differed significantly between experience groups within each TW. When a significant multivariate effect was observed, follow-up Welch's independent t-tests were performed to determine which specific AOI-based features contributed to the observed differences.

Results for learner versus expert groups:

- TW1: Pillai's Trace = 0.360, $F(20, 247) = 6.940, p = 0.000$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between learner and expert groups.
- TW2: Pillai's Trace = 0.231, $F(20, 247) = 2.330, p = 0.000$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between learner and expert groups.
- TW3: Pillai's Trace = 0.387, $F(20, 247) = 4.994, p = 0.000$. The multivariate effect was significant, indicating that eye movement patterns differed significantly between learner and expert groups.

To facilitate interpretation, the follow-up Welch's t-test results are presented in the subsequent sections, organized by feature category (Dwelling Time, Entry Count, and Entropy). Each section reports findings across all three TWs and summarizes the key AOI specific results.

8.3.1 Dwelling time

Pilots with different experience might exhibit different Dwelling time features on various AOIs across different TWs. The dwelling time features contain 'Airspeed_DT", "Attitude_DT", "Altitude_DT", "Roll_DT", "Heading_DT", "V Speed_DT", "Power_DT", "Front_DT" and "Others_DT", corresponding to 9 AOIs defined based on experiment scenarios. Each feature represents the percentage of time that pilots spent in the AOI.

Table 8.1 shows the results of Welch's independent t-tests between Dwelling time features of two experience groups. Figure 8.1 demonstrates the Dwelling time of all AOIs across three TWs, corresponding to Figure 8.1a, Figure 8.1b and Figure 8.1c. Each pair of bars represents the values from two different experience groups of pilots. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this Dwelling time feature between the two age groups. The error bars represent SEM of each bar.

Results During the first TW (60s before the alert; Figure 8.1a), expert pilots spent significantly more time focusing on the instrument panel than learner pilots. The instrument panel includes "Airspeed", "Attitude", "Altimeter", "Roll", "Heading" and "Vertical Speed" gauges. Expert pilots exhibited significantly longer dwelling times on all six gauges except "Airspeed". In contrast, expert pilots spent significantly less time in the "Front" and "Others" area than learner pilots, further supporting that their attention was concentrated on the instrument panel before the alert. This eye movement behavior aligns with expectations, since the aircraft was under autopilot control during TW1, and pilots should be primarily monitoring the instrument panel rather than looking outside or at other peripheral areas.

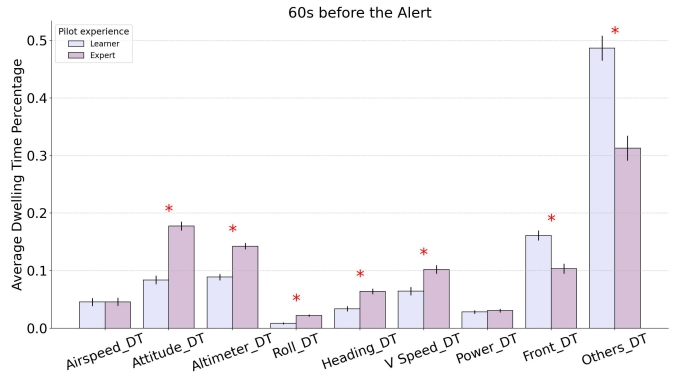
Between the alert and the takeover (TW2; Figure 8.1b), expert pilots showed significantly more Dwelling time only in the "Attitude" area, where the malfunction alert was presented. This suggests that expert pilots focused more on the source of the alert during this interval.

Table 8.1: T-tests for Dwelling time features between pilots from two experience groups across three TWs (* indicates $p < .05$)

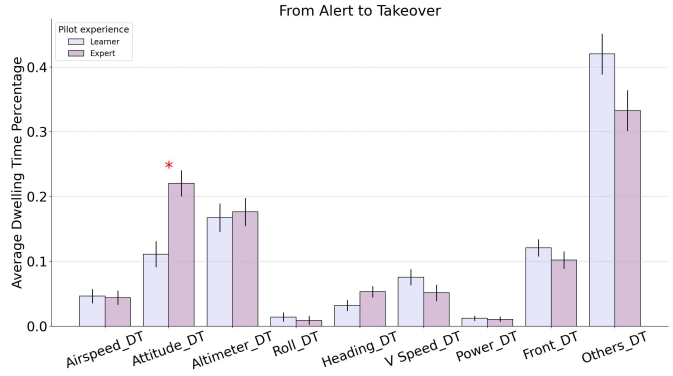
	Airspeed	**Attitude**	**Altimeter**	**Roll**	**Heading**	**V_Speed**	**Power**	**Front**	**Others**
TW1									
T-value	−0.022	−7.111	−5.272	−2.888	−4.110	−3.786	−0.632	5.114	6.396
P-value	0.983	**0.000***	**0.000***	**0.004***	**0.000***	**0.000***	0.528	**0.000***	**0.000***
Effect size	0.003	0.883	0.656	0.363	0.508	0.464	0.077	0.624	0.779
TW2									
T-value	0.155	−3.068	−0.264	0.606	−1.281	1.480	0.214	0.934	1.928
P-value	0.877	**0.002***	0.792	0.545	0.202	0.141	0.831	0.352	0.055
Effect size	0.023	0.468	0.040	0.091	0.196	0.223	0.033	0.142	0.292
TW3									
T-value	1.859	−4.621	−2.828	0.941	−2.634	−3.884	0.282	1.896	4.588
P-value	0.065	**0.000***	**0.005***	0.348	**0.009***	**0.000***	0.778	0.060	**0.000***
Effect size	0.275	0.696	0.431	0.140	0.403	0.594	0.042	0.287	0.675

Note: "V_Speed" refers to Vertical Speed. The "DT" in feature names is omitted to save space.

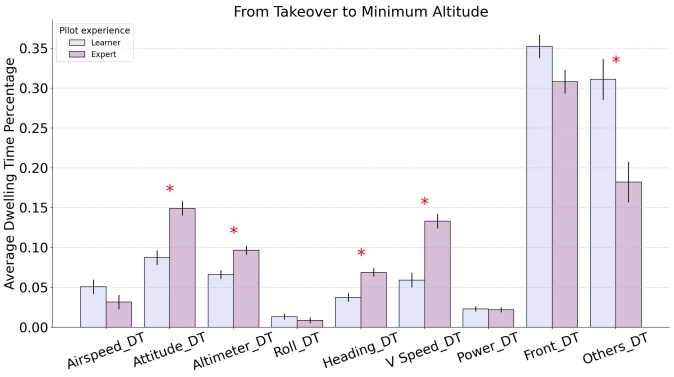
From the moment of takeover to the minimal altitude (TW3; Figure 8.1c), the pattern resembled that of TW1. Expert pilots had significantly more dwelling time in areas inside the instrument panel, including "Attitude", "Altimeter", "Heading" and "Vertical Speed", and significantly less time in the "Others" area. As pilots were manually flying the aircraft during this time, this pattern may reflect that expert pilots continued to prioritize monitoring the instrument panel over peripheral visual scanning, more so than the learner pilots.



(a) TW1



(b) TW2



(c) TW3

Figure 8.1: Dwelling time features of pilots from two experience groups across three TWs (error bars represent SEM).

8.3.2 Entry count

Pilots with different levels of experience may demonstrate distinct Entry count features on various AOIs across different TWs. The Entry count features contain "Airspeed_EC", "Attitude_EC", "Altitude_EC", "Roll_EC", "Heading_EC", "V Speed_EC", "Power_EC", "Front_EC" and "Others_EC", corresponding to 9 AOIs defined based on experiment scenarios. Each feature represents the percentage of time that pilots spent in the AOI.

Table 8.2 shows the result of Welch's indepdent t-tests between Entry count features of two TOT groups. Figure 8.2 demonstrates the Entry count of all AOIs across three TWs, corresponding to Figure 8.2a, Figure 8.2b and Figure 8.2c. Each pair of bars represents the values from two different experience groups of pilots. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this Entry count feature between the two age groups. The error bars represent SEM of each bar.

Table 8.2: T-tests for Entry count features between pilots from two experience groups across three TWs (* indicates $p < .05$)

	Airspeed	**Attitude**	**Altimeter**	**Roll**	**Heading**	**V_Speed**	**Power**	**Front**	**Others**
TW1									
T-value	0.685	−6.970	−4.177	−2.598	−4.016	−4.145	−0.858	4.451	3.829
P-value	0.494	**0.000***	**0.000***	**0.010***	**0.000***	**0.000***	0.392	**0.000***	**0.000***
Effect size	0.083	0.862	0.517	0.325	0.500	0.509	0.105	0.549	0.471
TW2									
T-value	1.901	−0.585	1.146	1.455	1.026	1.539	0.364	2.116	2.813
P-value	0.059	0.559	0.253	0.147	0.306	0.126	0.716	**0.036***	**0.005***
Effect size	0.283	0.089	0.175	0.218	0.153	0.231	0.056	0.320	0.420
TW3									
T-value	2.166	−3.876	−2.139	1.740	−2.137	−3.757	0.505	4.008	4.791
P-value	**0.032***	**0.000***	**0.034***	0.084	**0.034***	**0.000***	0.614	**0.000***	**0.000***
Effect size	0.320	0.584	0.326	0.255	0.327	0.571	0.075	0.601	0.713

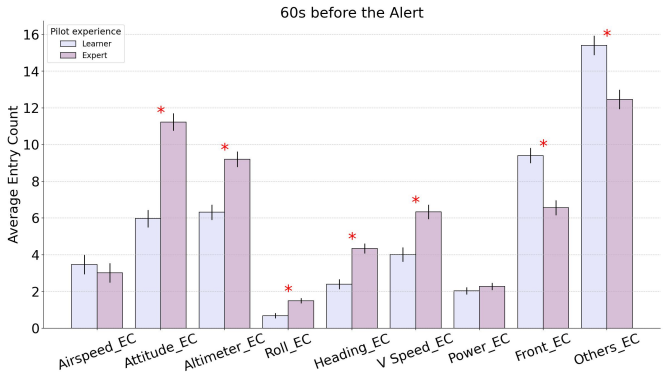
Note: "V_Speed" refers to Vertical Speed. The "EC" in feature names is omitted to save space.

Results During the first TW (60s before the alert; Figure 8.2a), expert pilots made significantly more entries into the gauges inside the instrument panel, including "Airspeed", "Attitude", "Altimeter", "Roll", "Heading" and "Vertical Speed", and fewer entries into the "Front" and "Others" areas. These significant differences are similar to the findings

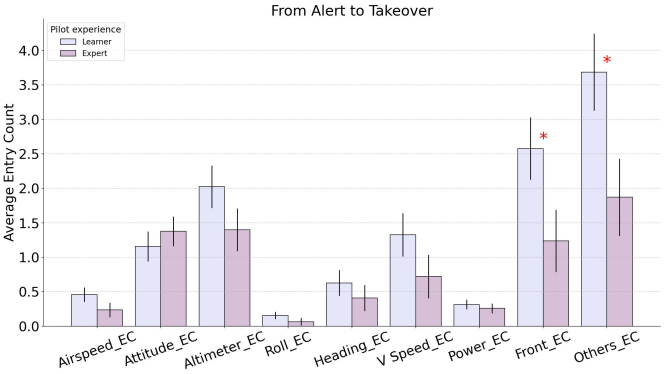
from the Dwelling time results in TW1, further supporting that expert pilots engaged in more active scanning within the instrument panel and paid less attention to outside or peripheral regions during this time.

After the alert and before the takeover (TW2; Figure 8.2b), expert pilots only had significantly fewer entries into the "Front" and "Others" areas, which implies that they checked the outside and surrounding areas with less frequently after the alert was issued.

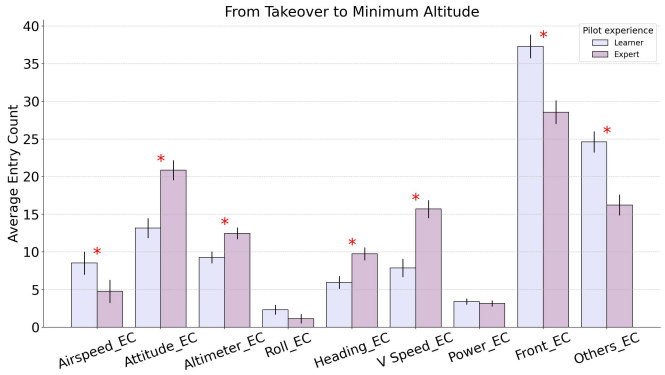
From the moment of takeover to the minimal altitude (TW3; Figure 8.2c), the eye movement pattern of expert pilots reverted to a trend similar to that of TW1, where they had significantly more Entry count in areas inside the instrument panel, including "Airspeed", "Attitude", "Altimeter", "Heading" and "Vertical Speed", and significantly fewer Entry count in "Front" and "Others". This suggests that during manual flight, expert pilots continued to perform more frequent visual scanning of instruments and less monitoring of the external environment compared to learner pilots.



(a) TW1



(b) TW2



(c) TW3

Figure 8.2: Entry count features of pilots from two experience groups across three TWs (error bars represent SEM).

8.3.3 Entropy

Pilots with different levels of experience may demonstrate distinct SGE and GTE across different TWs.

Table 8.3 shows the result of Welch's independent t-tests between SGE and GTE of two TOT groups. Figure 8.3 demonstrates the SGE and GTE of all AOIs across three TWs, corresponding to Figure 8.3a, Figure 8.3b and Figure 8.3c. Each pair of bars represents the values from two different experience groups of pilots. Independent-sample t-tests were conducted for each pair, and a red asterisk was placed above the bars where a statistically significant difference was found. In such cases, this indicates a significant difference in this entropy feature between the two age groups. The error bars represent SEM of each bar.

Table 8.3: T-tests for SGE and GTE between pilots from two experience groups across three TWs (* indicates $p < .05$)

	SGE	**GTE**
TW1		
T-value	−2.833	−2.185
P-value	**0.005***	**0.030***
Effect size	0.347	0.266
TW2		
T-value	3.458	3.265
P-value	**0.001***	**0.001***
Effect size	0.524	0.488
TW3		
T-value	−0.619	0.537
P-value	0.081	0.935
Effect size	0.092	0.012

Results During the first TW (60s before the alert; Figure 8.3a), expert pilots showed significantly higher SGE and GTE compared to learner pilots. This suggests that expert pilots had a more dispersed gaze distribution and engaged in more complex transitions across different AOIs, indicating a more active and varied scanning strategy.

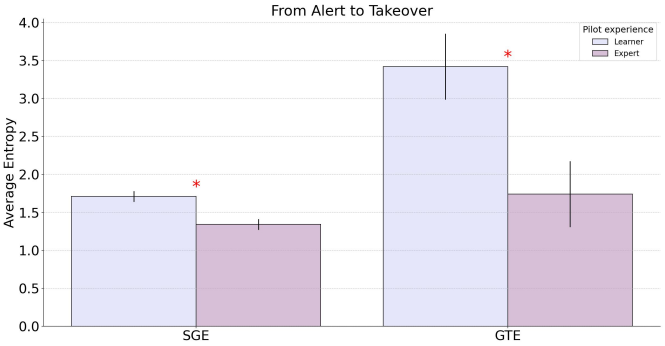
After the alert and before the takeover (TW2; Figure 8.3b), expert pilots demonstrated significantly lower SGE and GTE compared to learner pilots. This indicates that expert

pilots employed a more focused gaze distribution and more predictable transitions between AOIs, possibly concentrating on specific areas relevant to the alert.

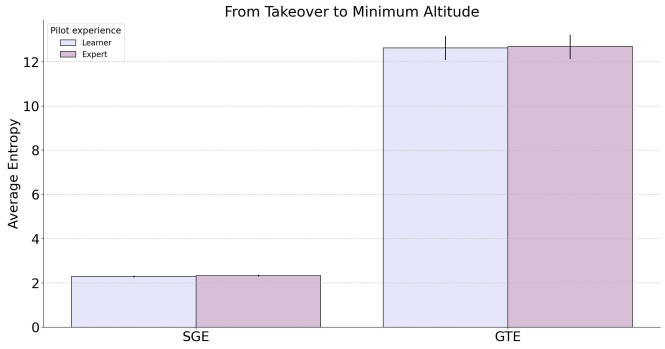
From the moment of takeover to the minimal altitude (TW3; Figure 8.3c), no significant differences in either SGE or GTE were observed between the two experience groups.



(a) TW1



(b) TW2



(c) TW3

Figure 8.3: SGE and GTE of pilots from two experience groups across three TWs (error bars represent SEM).

8.4 Discussion and conclusions

This study examined experience-related differences in pilots' eye movement behavior during the takeover process in an autopilot-assisted landing, focusing on four categories of eye-tracking features: Dwelling Time, Entry count, SGE, and GTE. The results consistently showed that expert pilots exhibited distinct visual scanning patterns compared to leaner pilots across different TWs of the takeover process.

During the pre-alert phase (TW1), expert pilots showed significantly more Dwelling time and higher Entry counts in instrument panel areas, along with significantly higher values in both SGE and GTE. These results suggest that expert pilots engaged in more thorough and diverse scanning of the instrument, likely reflecting a proactive monitoring strategy while the aircraft was under autopilot control. Moreover, their scanning behavior was more dispersed and followed a more complex sequential pattern.

After the alert and before the takeover (TW2), expert pilots focused more narrowly on the task-relevant area. They exhibited significantly more Dwelling time and significantly higher Entry counts in the "Attitude" area —where the malfunction alert appeared — while allocating significantly less attention to peripheral areas such as the "Front" and "'Others". Correspondingly, they exhibited significantly lower SGE and GTE values, suggesting a more focused and predictable visual strategy during this critical reaction period.

During the manual control phase (TW3), the pattern largely reverted to what was observed in TW1. Expert pilots again showed significantly greater Dwelling time and Entry counts on key instruments, and fewer entries into peripheral areas. No significant differences in entropy values were found between the two groups during this phase, suggesting that both groups may have adopted similarly focused visual strategies during manual flight.

Together, these findings demonstrate that expert pilots not only allocate more attention to critical instrument areas but also dynamically adjust their scanning behavior based on the current tasks. They demonstrated a clear understanding of the ongoing task demands, whether it involved monitoring flight instruments or promptly identifying system malfunctions. These findings offer valuable insights for the development of gaze training programs aimed at helping learner pilots develop more efficient visual attention strategies.

Chapter 9

Using eye tracking features to predict drivers' Situation Awareness and takeover performance in a semi-autonomous driving system with machine learning methods.

This chapter presents the driving domain results and discussion of Topic 3: **Using eye tracking features to predict operator's Situation Awareness and takeover performance in a conditionally automated system with machine learning methods.**

9.1 Background

In conditionally automated driving systems, the transition from automated to manual control—commonly referred to as the takeover process—could be challenging to drivers. It is likely that drivers may exhibit inadequate SA due to the engagement of NDRTs, thereby leading to poor takeover performance, characterized by delayed or inappropriate control actions. Extensive research has shown that eye movement behavior is closely linked to drivers' SA and performance. Therefore, monitoring eye movement offers a promising, non-invasive method for assessing drivers' behaviors. While the current work focuses on offline model development, it lays the groundwork for future integration into real-time monitoring

systems. By developing an automated approach to evaluate drivers' SA and takeover performance based on their eye-tracking data, this study aims to eventually contribute to proactive safety mechanisms in semi-autonomous systems. Such an approach could enable real-time monitoring and adaptive interventions, ultimately improving road safety by supporting drivers during critical handover scenarios.

9.2 Methodology

In the current study, one driving experiment using a level-3 autonomous vehicle simulator was completed as introduced in Chapter 3.1. Drivers' eye movement features and AOI features were extracted as discussed in Chapter 4.1 and Chapter 4.1.3.1, and they are the input of the machine learning models. A series of machine learning models — outlined in the pipeline presented in Chapter 4.4.2.2 — were developed to predict drivers' SA and takeover performance. Drivers' SA were measured by their SAGAT score and HPT, and their takeover performance was measured by their TOT and Resulting accelerations. Each of these four metrics were grouped into binary values (see Chapter 4.4.1), and accordingly, therefore all predictive models were formulated as classification models.

For each classification model, a grid search was conducted to identify the optimal hyper-parameters. Once the best hyper-parameters were selected, each model was retrained and evaluated using 5-fold stratified cross-validation. To comprehensively assess model performance, a variety of evaluation metrics were computed, including accuracy, precision, recall, F1 score, and AUC. Although confusion matrices were generated for all models to visualize classification performance across folds, only the confusion matrix corresponding to the best-performing model was included here.

There were 12 eye movement features related to the fixation and saccades were included, along with four types of AOI feature: Dwelling time, Entry count, SGE and GTE. After processing the data, we found that the dwelling time and entry counts of three mirrors: Rear-mirror, Left mirror, and Right mirror were relatively low. Therefore, these three AOIs were combined into one single AOI labeled as the "Mirrors". Therefore, we have 4 Dwelling Time and 5 Entry Count features. Additionally, two entropy features — SGE and GTE — were included to quantify uncertainty in eye movement. In total, the input variables of the driving study are based on the following 23 eye-tracking features: 12 eye movement features, 11 AOI features (4 Dwelling Time + 5 Entry Count), and 2 entropy features. These 26 features were extracted for each relevant TW and concatenated to form the input to the machine learning model. Therefore, if the predicted metric was relevant to all three TWs (e.g. SAGAT score), there were 23*3 input variables. If the predicted

metric was only relevant to a single TW (e.g. TOT was only relevant to the first TW, "Pre-TOR"), there were only 26 input variables.

Five metrics were predicted in the current driving study: SAGAT score, HPT, TOT and Resulting acceleration. Separate models were developed for each of the five metrics, resulting in five independent classification tasks.

9.3 Results

9.3.1 Situation Awareness

9.3.1.1 Drivers' SAGAT score prediction

Eye tracking features served as the input for the machine learning models and were extracted for each TW described earlier. For the SAGAT score prediction, all TWs are considered relevant since the SAGAT questions involved information spanning the entire trial. Therefore, features from all three TWs — "10s before the TOR", "from TOR to hazard perceived" and "from hazard perceived to the end" — were used as input to the classification models. The classification task involved two target classes with a near-balanced distribution: 51% for the higher SAGAT score and 49% for the lower SAGAT score.

Table 9.1: Classification performance of five models on the drivers' SAGAT score

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	58.85%	57.89%	57.54%	57.48%	59.82%
Linear SVM	59.06%	58.11%	57.73%	57.67%	60.08%
RBF SVM	**59.70%**	**58.85%**	**58.56%**	**58.56%**	**61.33%**
Random Forest	59.06%	58.07%	56.67%	55.82%	57.56%
Gaussian Process	52.45%	51.48%	51.46%	51.43%	43.63%

Table 9.1 presents the classification performance of five machine learning models in distinguishing between two SAGAT score groups. Among the models tested, the RBF SVM achieved the highest overall performance, with an accuracy of 59.70%, precision of 58.85%, recall of 58.56%, F1 score of 58.56%, and AUC of 61.33%. While this performance is modest, it consistently outperformed the other models across all evaluation metrics.

Figure 9.1 presents the confusion matrix of the RBF SVM model in predicting binary SAGAT scores (higher versus lower). The model correctly classified 179 participants with high SAGAT scores (true positives) and 101 participants with low SAGAT scores (true negatives), while misclassifying 83 high-SAGAT individuals as low (false negatives) and 106 low-SAGAT individuals as high (false positives). In all, the model demonstrated a slightly better ability to correctly identify high-SAGAT individuals, albeit with a relatively high false positive rate. Overall, while the RBF SVM shows moderate performance, the substantial number of misclassifications suggests that the eye movement features used may have limited predictive power for distinguishing between higher and lower SAGAT scores.

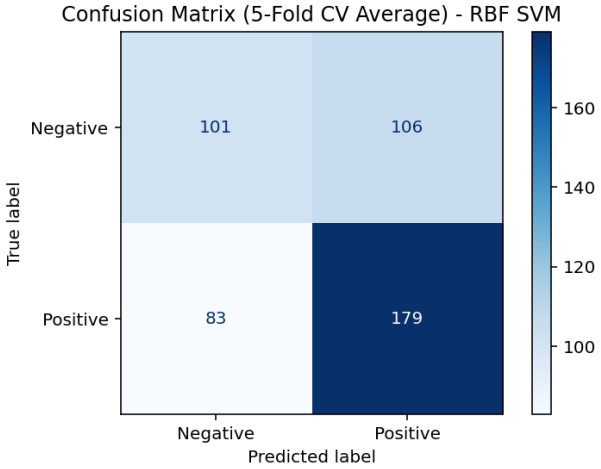


Figure 9.1: Confusion matrix of the RBF SVM model in classifying higher vs. lower SAGAT scores

9.3.1.2 Drivers' Hazard perception prediction

Eye tracking features served as the input for the machine learning models and were extracted for each TW described earlier. For the HPT prediction, only features from the two TWs— "10s before the TOR", "from TOR to hazard perceived" — were considered relevant. Therefore, only eye tracking features from these two TWs were used as input to the classification models. The classification task involved two target classes with a moderately imbalanced distribution: 29% for the longer HPT and 71% for the shorter HPT.

Table 9.2 presents the classification performance of five machine learning models in distinguishing between drivers with shorter versus longer Hazard Perception Time. Among

Table 9.2: Classification performance of five models on the drivers' Hazard Perception time

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	83.51%	80.25%	75.92%	77.60%	87.03%
Linear SVM	83.72%	80.22%	76.80%	78.20%	86.20%
RBF SVM	83.30%	80.32%	75.04%	76.97%	87.65%
Random Forest	77.17%	72.90%	62.41%	63.86%	81.06%
Gaussian Process	**84.36%**	**81.50%**	**76.99%**	**78.75%**	**88.48%**

all models evaluated, the Gaussian Process classifier achieved the highest performance, with an accuracy of 84.36%, precision of 81.50%, recall of 76.99%, F1 score of 78.75%, and an AUC of 88.48%. These results indicate strong and consistent predictive capability of Gaussian Process classifier across all evaluation metrics.

Figure 9.2 shows the confusion matrix for the Gaussian Process classifier, averaged over five-fold cross-validation. The model correctly classified 321 drivers with shorter HPT (true negatives) and 78 drivers with longer HPT (true positives), while misclassifying 24 short-HPT drivers as longer HPT (false positives) and 50 longer HPT drivers as shorter HPT(false negatives). The relatively balanced classification, with higher sensitivity to short-HPT individuals, demonstrates the model's robustness and suggests that the selected eye movement features carry meaningful predictive value for differentiating drivers' HPT.

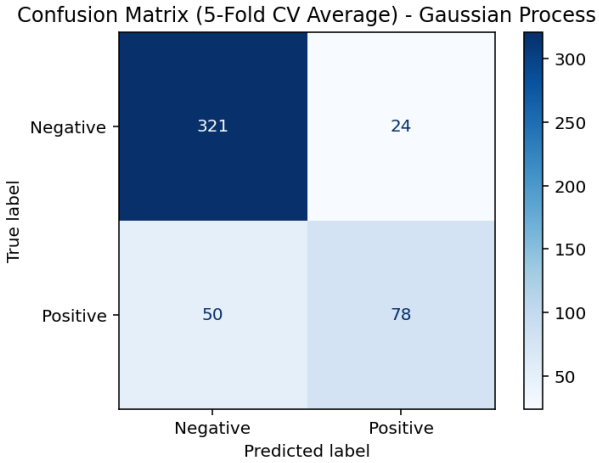
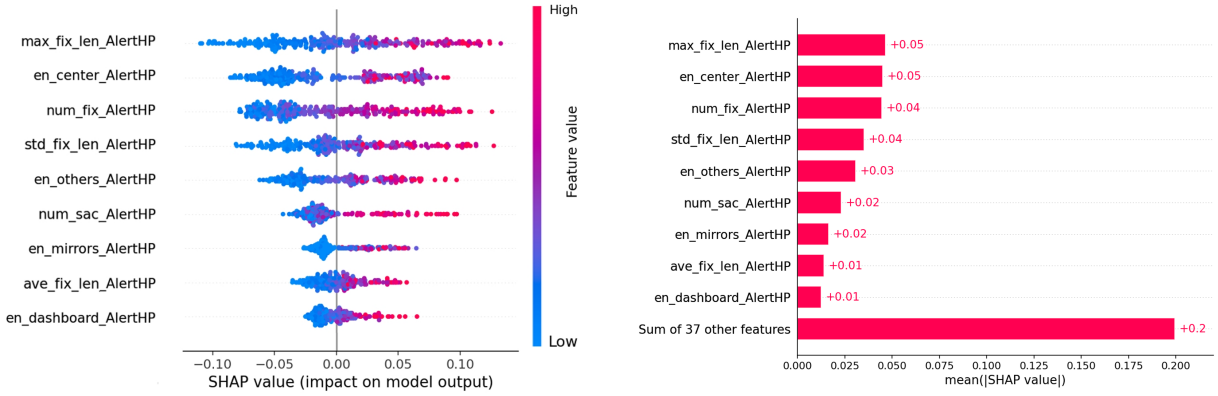


Figure 9.2: Confusion matrix of the Gaussian Process model in classifying shorter vs. longer hazard perception time

To further interpret the eye tracking features, SHAP analysis was conducted on the GPC model above. As shown in Figure 9.3, the top contributing features included *max_fix_len_AlertHP* (maximun fixation length in TW2), *en_center_AlertHP* (Entry counts to the "Center" AOI in TW2), and *num_fix_AlertHP* (number of fixations in TW2), with average absolute SHAP values of approximately 0.05, 0.05, and 0.04, respectively. These results suggest that eye tracking features from the TW2 (from TOR to hazard perceived) played a key role in predicting drivers' HPT. The feature rankings further indicated that a higher maximum fixation length and a greater number of fixations are associated with longer HPT predictions. Additionally, more entries to the "Center" AOI slightly push predictions toward longer HPT, which aligns with the results in Chapter 5, where drivers with shorter HPT exhibited significantly less Entry counts in multiple AOIs during TW2.



(a) SHAP value distribution.

(b) Mean absolute SHAP values.

Figure 9.3: SHAP values ranking of the top 9 eye tracking features used in the HPT classification task.

9.3.2 Takeover performance

9.3.2.1 Drivers' Takeover time prediction

Eye tracking features served as the input for the machine learning models and were extracted for each TW described earlier. For the TOT prediction, only features from "10s before the TOR" TW were considered relevant. Therefore, only eye tracking features from this TW were used as input to the classification models. The classification task involved two target classes with a moderately imbalanced distribution: 35% for the longer TOT and 65% for the shorter TOT.

Table 9.3: Classification performance of five models on the drivers' Takeover time

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	63.18%	54.88%	52.17%	48.71%	54.50%
Linear SVM	64.19%	48.92%	49.96%	39.98%	44.83%
RBF SVM	**65.88%**	**61.54%**	**59.62%**	**59.81%**	**63.46%**
Random Forest	65.54%	60.53%	55.93%	54.47%	62.98%
Gaussian Process	65.20%	70.03%	51.17%	42.09%	56.47%

Table 9.3 presents the classification performance of five machine learning models in distinguishing between drivers with shorter versus longer Takeover Time. Among the models evaluated, the RBF SVM achieved the best overall performance, with an accuracy of 65.88%, precision of 61.54%, recall of 59.62%, F1 score of 59.81%, and AUC of 63.46%. This model consistently outperformed the others across most metrics, indicating a relatively strong ability to capture eye movement patterns associated with takeover speed.

Figure 9.4 shows the confusion matrix of the Gaussian Process model, which was the most precise classifier, achieving a precision of 70.03%. While the Gaussian Process correctly identified 321 shorter TOT drivers and 78 longer TOT drivers, it also misclassified 50 longer TOT individuals as shorter TOT (false negatives) and 24 shorter TOT individuals as longer TOT (false positives). Although the Gaussian Process model demonstrated high precision, its relatively low F1 score suggests it struggled with balanced classification performance. Overall, the results suggest that the RBF SVM may offer the most balanced prediction capability for TOT, while the Gaussian Process model may be more conservative in flagging longer takeover times.

9.3.2.2 Drivers' Resulting acceleration prediction

Eye tracking features served as the input for the machine learning models and were extracted for each TW described earlier. For the Resulting acceleration prediction, all TWs are considered relevant since this metric was calculated from the whole trial. Therefore, features from all three TWs — "10s before the TOR", "from TOR to hazard perceived" and "from hazard perceived to the end" — were used as input to the classification models. The classification task involved two target classes with a severe imbalanced distribution: 5% for the higher Resulting acceleration and 95% for the lower Resulting acceleration.

Table 9.4 presents the classification performance of five machine learning models in predicting drivers' resulting acceleration. All models achieved exceptionally high accuracy

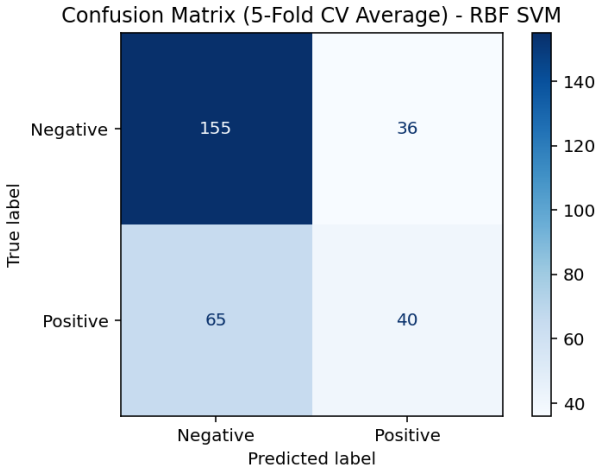


Figure 9.4: Confusion matrix of the RBF SVM model in classifying shorter vs. longer Takeover time

Table 9.4: Classification performance of five models on the resulting acceleration

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
Logistic Regression	87.64%	49.55%	49.70%	49.47%	60.51%
Linear SVM	86.18%	51.08%	50.99%	51.02%	57.38%
RBF SVM	92.00%	46.00%	50.00%	47.92%	37.46%
Random Forest	92.00%	46.00%	50.00%	47.92%	52.71%
Gaussian Process	92.00%	46.00%	50.00%	47.92%	44.77%

scores—up to 92.00%—with the RBF SVM, Random Forest, and Gaussian Process models performing identically across all metrics. However, this high accuracy is misleading, as the models yielded relatively low precision (46.00%), recall (50.00%), and F1 scores (47.92%), along with modest AUC values. These results suggest that the models may have been affected by a significant class imbalance, likely favoring the majority class during training and evaluation. In such cases, accuracy becomes a less reliable indicator of true model performance, as it can be dominated by correct predictions of the majority class. Future work should consider employing resampling techniques (e.g., SMOTE), adjusting class weights, or using more balanced datasets to improve classification fairness and performance on this metric.

9.4 Discussion and conclusions

This study investigated the feasibility of using eye tracking features to predict drivers' cognitive and behavioral outcomes during the takeover process. Across the four classification tasks—predicting drivers' binary SAGAT scores, HPT, TOT, and Resulting acceleration — the performance of the machine learning models varied depending on the target metric.

For HPT, the Gaussian Process model delivered the best performance across all metrics, including accuracy (84.36%), F1 score (78.75%), and AUC (88.48%). This strong performance suggests that binary HPT is a relatively well-structured and learnable outcome, indicating a stronger association between drivers' eye movement and their hazard anticipation. Gaussian Process classifiers, known for their probabilistic modeling and capacity to handle subtle nonlinearities, were well suited for this task.

In contrast, for SAGAT, which involves higher-level cognitive processes captured through the SAGAT questionnaire, the RBF SVM model outperformed others. Although its performance was modest (accuracy: 59.70%, F1: 58.56%, AUC: 61.33%), it consistently led across all evaluation metrics. This suggests that when using eye tracking features to predict drivers' SA, the produced feature exhibit non-distinct patterns, due to the abstract and noisy nature of drivers' SA. As a result, the distributions of the two SA groups substantially overlap, making linear models less effective. In contrast, kernel-based SVMs (such as RBF SVM) are better equipped to handle such nonlinearly separable and noisy data by learning more flexible decision boundaries.

For TOT, the RBF SVM again achieved the best performance (accuracy: 65.88%, F1: 59.81%), indicating that the prediction of takeover time shares a similar feature structure to SAGAT scores, where flexible decision boundaries is required. The Gaussian Process model, while highly precise, underperformed on recall and F1, suggesting it was overly conservative in classifying longer TOT cases.

Finally, the Resulting acceleration results revealed performance limitations across all models used. Although several models achieved very high accuracy (up to 92.00%), their precision, recall, F1 score, and AUC remained low, revealing an underlying class imbalance issue. The models likely over-predicted the dominant class, masking true performance. This indicates that resulting acceleration may be a harder-to-learn target using eye movement features alone, and requires additional pre-processing (e.g., class balancing or feature engineering) to improve model robustness.

In summary, Gaussian Process models performed best when the target was more structured and measurable (e.g., HPT), while SVMs with RBF kernels were more effective

for cognitively complex or noisier outcomes (e.g., SAGAT score and TOT). The performance drop in biACC highlights the importance of accounting for data imbalance and target-specific variability when applying machine learning to behavioral prediction tasks in driving contexts.

Chapter 10

Using eye tracking features to predict pilots' Situation Awareness and takeover performance in an autopilot-assisted instrument landing with machine learning methods.

This chapter presents the piloting domain results and discussion of Topic 3: **Using eye tracking features to predict operator's Situation Awareness and takeover performance in a conditionally automated system with machine learning methods.**

10.1 Background

In current pilot training and assessment systems, pilot performance is typically evaluated by flight instructors based on their subjective judgment. As the global demand for trained pilots continues to rise, there is an urgent need for scalable, objective, and consistent methods of pilots' performance evaluation. One promising direction is the development of automated assessment tools. In the current study, a pilot's SA and takeover performance during transitions from autopilot to manual control are critical to flight safety. Pilots' eye movement offers a real-time, non-intrusive window into how pilots allocate visual attention and process flight-critical information. Building on this foundation, the present study

explores the use of machine learning models to predict pilots' SA and takeover performance using eye-tracking features. This approach has the potential to offer a data-driven pathway toward more efficient, standardized, and scalable pilot assessments.

10.2 Methodology

In the current study, one piloting experiment using ILS was completed as discussed in Chapter 3.2. Pilots' eye movement features and AOI features were extracted as discussed in Chapter 4.1. A series of machine learning models — outlined in the pipeline presented in Chapter 4.4.2.2 — were developed to predict pilots' SA, which were measured by their SAGAT score, and their takeover performance, which was measured by their TOT. Each of these two metrics were grouped into binary values (see Chapter 4.4.1), and accordingly, all predictive models were formulated as binary classification models.

For each classification model, a grid search was conducted to identify the optimal hyper-parameters. Once the best hyper-parameters were selected, each model was retrained and evaluated using 5-fold stratified cross-validation. Before model training, all features went through standard scaler and PCA (95% original variance preserved). To comprehensively assess model performance, a variety of evaluation metrics were computed, including accuracy, precision, recall, F1 score, and AUC. Although confusion matrices were generated for all models to visualize classification performance across folds, only the confusion matrix corresponding to the best-performing model was included here.

The input to the machine learning models consisted of eye tracking features extracted following the steps in Section 4.1. Two categories of eye tracking features were included: eye movement feature and AOI feature. There were 12 eye movement features related to the fixation and saccades were included, along with four types of AOI feature: Dwelling time, Entry count, SGE and GTE. Dwelling Time and Entry Count were computed separately for each of the nine pre-defined AOIs in the piloting study, resulting in 9 Dwelling Time and 9 Entry Count features. Additionally, two entropy features — SGE and GTE — were included to quantify uncertainty in eye movement.

In total, the piloting study used 32 eye-tracking features: 12 eye movement features, 18 AOI features (9 Dwelling Time + 9 Entry Count), and 2 entropy features. These 32 features were extracted for each relevant TW and concatenated to form the input to the machine learning model. Therefore, if the predicted metric was relevant to all three TWs (e.g. SAGAT score), there were 32*3 input variables. If the predicted metric was only relevant to a single TW (e.g. TOT was only relevant to the first TW, "Pre-alert"), there

were only 32 input variables. The output of the models consisted of two metrics being predicted from the current piloting study: SAGAT scores, TOT. Separate models were developed for each of the three metrics, resulting in three independent classification tasks.

10.3 Results

10.3.1 Situation Awareness

10.3.1.1 Pilots' SAGAT score prediction

Eye tracking features served as the input for the machine learning models and were extracted for each TW described earlier. For the SAGAT score prediction, all TWs are considered relevant since the SAGAT questions involved information spanning the entire trial. Therefore, features from all three TWs — "60s before the alert", "from alert to Takeover" and "from Takeover to Minimum Altitude" — were used as input to the classification models. The classification task involved two target classes with a moderately balanced distribution: 66% for the higher SAGAT score and 34% for the lower SAGAT score.

Table 10.1: Classification performance of five models on the pilots' SAGAT scores

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	66.67%	60.80%	58.62%	58.81%	58.90%
Linear SVM	62.64%	56.93%	56.47%	56.58%	50.61%
RBF SVM	63.79%	57.49%	56.47%	56.56%	42.97%
Random Forest	**68.39%**	**63.15%**	**57.75%**	**57.19%**	**53.64%**
Gaussian Process	64.37%	55.68%	53.45%	52.12%	55.63%

Table 10.1 presents the classification performance of five machine learning models in distinguishing between two SAGAT score groups. Among the models evaluated, the Random Forest classifier achieved the highest overall performance, with an accuracy of 68.39%, precision of 63.15%, recall of 57.75%, F1 score of 57.19%, and an AUC of 53.64%. These results, although moderate, suggest that the Random Forest model was better able to capture the complex patterns in pilots' eye movement data that are associated with SA-related cognitive processing, compared to the other classifiers.

Figure 10.1 illustrates the confusion matrix for the Random Forest model. The model correctly identified 104 pilots with higher SAGAT scores (true positives) and 15 pilots with lower SAGAT scores (true negatives), while misclassifying 12 pilots with higher SAGAT scores as low (false negatives) and 43 pilots with lower SAGAT scores as high (false positives). The model exhibited a stronger capability in detecting higher-SAGAT scores individuals, though at the cost of a relatively high false positive rate.

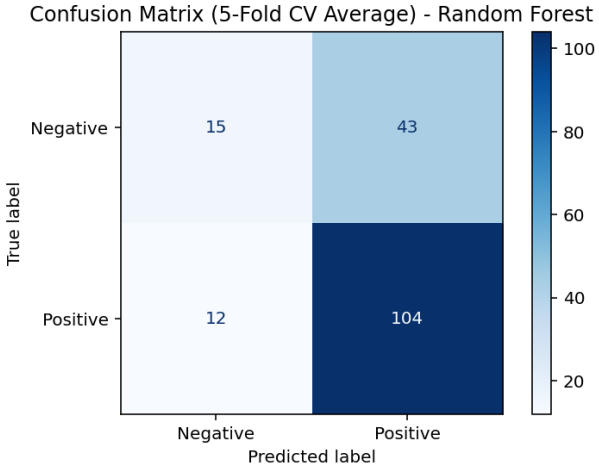


Figure 10.1: Confusion matrix of the Random Forest model in classifying higher vs. lower SAGAT scores

10.3.2 Takeover performance

10.3.2.1 Pilots' Takeover time prediction

Eye tracking features served as the input for the machine learning models and were extracted for each TW described earlier. For the TOT prediction, only features from "60s before the alert" and "from alert to Takeover" two TWs were considered relevant. Therefore, only eye tracking features from these two TWs were used as input to the classification models. The classification task involved two target classes with a severe imbalanced distribution: 13% for the longer TOT and 87% for the shorter TOT.

Table 10.2 presents the classification performance of five machine learning models in classifying between pilots' shorter versus longer TOT. Among the models, the Random Forest classifier achieved the highest overall performance, with an accuracy of 87.50%,

Table 10.2: Classification performance of five models on the pilots' takeover time

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	80.68%	50.00%	50.00%	49.85%	**65.32%**
Linear SVM	81.25%	47.44%	48.38%	47.65%	52.27%
RBF SVM	82.95%	48.67%	49.36%	48.44%	56.82%
Random Forest	**87.50%**	**68.96%**	**51.94%**	**58.13%**	54.94%
Gaussian Process	87.50%	43.75%	50.00%	46.67%	53.45%

precision of 68.96%, recall of 51.94%, and F1 score of 58.13%, showing its ability to capture important reaction speed from pilots' eye movement features. While Logistic Regression achieved the highest AUC (65.32%), its lower precision and F1 score suggest that it was less effective at classifying both classes simultaneously.

Figure 10.2 displays the confusion matrix for the Random Forest model. The model correctly identified 153 pilots with shorter TOT (true negatives) and only one pilot with a longer takeover time (true positive), while misclassifying 21 pilots with longer TOT as short (false negatives) and 1 pilot with shorter TOT as long-TOT (false positive). These results reveal a strong bias toward the majority class, leading to high overall accuracy but poor sensitivity to longer takeover time cases. Despite the model's high accuracy, its limited generalizability to the minority class reflects a class imbalance issue. This limitation should be considered when interpreting the results, and future efforts may benefit from balancing strategies or alternative metrics better suited for imbalanced classification.

10.4 Discussion and conclusions

This study investigated the feasibility of using eye movement features to predict pilots' cognitive and behavioral outcomes during the takeover process from automated flight control. Two different variables were examined: pilots' SA as measured by the SAGAT score, and pilots' takeover performance as measured by their TOT. Across these two prediction tasks, it is concluded that the success of machine learning models in each task was closely linked to the nature of the outcome variable.

For predicting pilots' SA, the Random Forest model yielded the highest performance (accuracy: 68.39%, F1: 57.19%). While this was modest, it outperformed other classifiers and demonstrated a relatively balanced ability to distinguish between higher and lower SA levels. However, the model tended to favor higher SAGAT score predictions, as evidenced

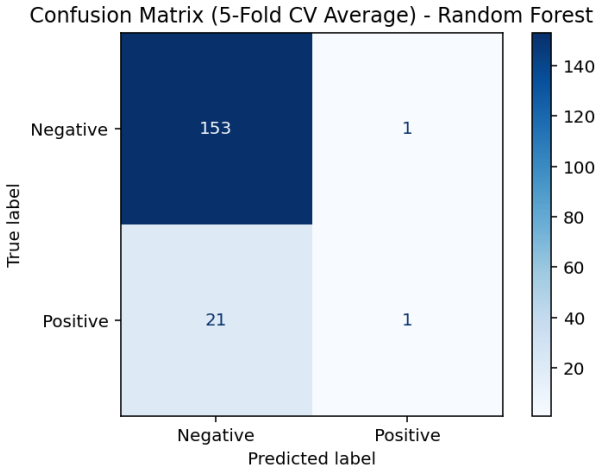


Figure 10.2: Confusion matrix of the Random Forest model in classifying shorter vs. longer takeover time

by a higher false positive rate, suggesting that SA is difficult to infer solely from eye tracking data.

In the prediction of TOT, or pilots' takeover time, Random Forest again delivered the best overall performance with an accuracy of 87.50% and an F1 score of 58.13%. However, the confusion matrix of the Random Forest model revealed that the model was heavily biased toward predicting shorter TOT, with almost no correct identification of longer takeover time cases. This indicates a class imbalance issue, where the model's high accuracy masked a failure to generalize across both classes.

The comparatively better performance performance of the Random Forest models in this study suggests that the underlying structure of the eye-tracking dataset remained highly nonlinear, even after dimensionality reduction using Principal Component Analysis (PCA). Random Forest models remained effective in identifying meaningful patterns within the transformed space. Their ability to partition the feature space nonlinearly allowed them to outperform linear models (e.g., Logistic Regression, Linear SVM) that are more constrained in the types of relationships they can capture in PCA-transformed data.

Chapter 11

Synthesis, Limitations, and Future Work

After exploring three distinct but interrelated topics in Chapter 5 to 10, this final chapter integrates additional findings by comparing results across topics as well as domains. It also discusses the thesis limitations, outlines future research directions, and clarifies its contributions to scientific knowledge, methodology, design, and training.

11.1 Synthesis between topics and domains

While three topics focused on a specific aspect, a comparative lens allows us to uncover broader trends and identify domain-specific distinctions. While not all findings led to integrated or cross-domain conclusions, a subset revealed consistent patterns that support broader insights into operator eye movement. Overall, those insights highlight the importance of accounting for task goal across different TWs when examining operators' eye movement. Given the experimental scenarios, operators' goals varied in each TW, leading to corresponding changes in optimal eye movement patterns.

The conclusions were derived through two sets of comparisons: between Topic 1 and Topic 2 (Comparison A), and between the driving and piloting domains (Comparison B). Figure 11.1 illustrates the the structure of these comparisons, where only the relevant previous findings are included. Comparisons A and B are discussed in the following sections.

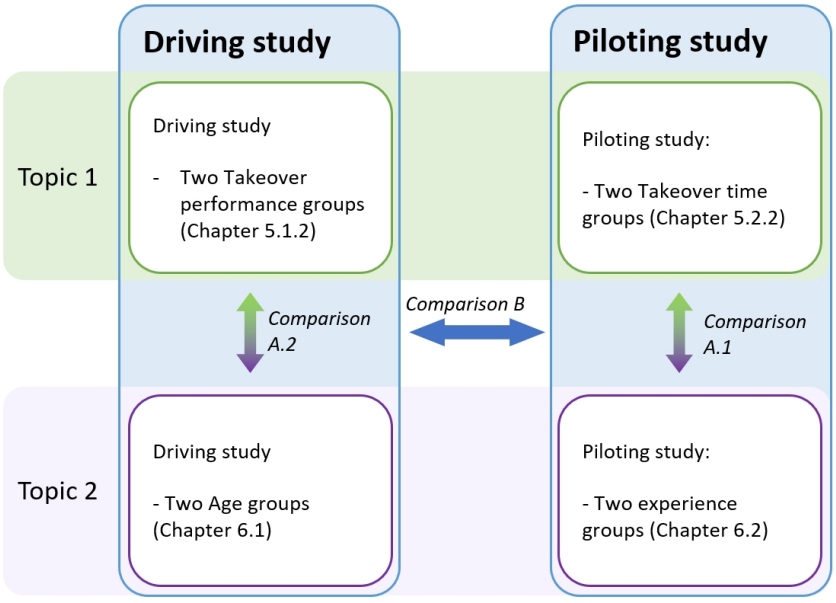


Figure 11.1: Comparisons between results from two topics and two domains

11.1.1 Comparison between Topic 1 and 2

11.1.1.1 The piloting study (Comparison A.1)

In the piloting domain, three different TWs were identified, as introduced in Chapter 4.3.2 — Pre-alert (TW1), from alert to Takeover (TW2) and from Takeover to Minimum Altitude (TW3). During TW1, the pilot's task was to monitor the autopilot system. In TW2, the task shifted to detecting the alert and determining the appropriate response, which is disengaging the autopilot. In TW3, the pilot was responsible for manually controlling the aircraft and ensuring a safe landing.

In Topic 1, pilots' eye movements were compared between shorter-TOT and longer-TOT groups, while in Topic 2, the same eye tracking features were examined between learner and expert pilots. Similar eye movement patterns were observed between shorter-TOT pilots and expert pilots groups, across three TWs. As discussed in Chapter 6.3.2.2, shorter-TOT pilots showed distinct patterns in Entry count across the three TWs. In TW1, they made significantly more entries into the "Attitude" area. In TW2, they showed significantly fewer entries across almost all AOIs. In TW3, their entry count increased again, particularly in four AOIs, including two within the instrument panel as well as the "Front" and "Others" areas. Regarding entropy, shorter-TOT pilots exhibited higher SGE and GTE in TW1, and

higher GTE in TW3, but lower SGE and GTE in TW2 (see Chapter 6.3.2.3). Expert pilots showed similar eye movement behaviors. As noted in Chapter 8.3.2, they had higher Entry counts in five AOIs inside the instrument panel and lower Entry counts in the "Front" and "Others" areas during TW1. In TW2, their entries in "Front" and "Others" were still significantly less. In TW3, their gaze distribution largely returned to the same pattern as in TW1. Their entropy values followed a similar trend — SGE and GTE are both higher in TW1; SGE and GTE are both lower in TW2 (see Chapter 8.3.3).

Although shorter TOT and higher experience (expert pilots) reflect different aspects, they both are considered as desirable characteristics in the current thesis. Accordingly, they are collectively referred to as the "better" group. When pilots are engaged in routine monitoring or standard manual control tasks (e.g. TW1 and TW3), the better group would regularly scan multiple AOIs to gather necessary information. In such cases, pilots in the better groups typically exhibit significantly higher Entry counts across multiple AOIs and higher entropy values. Higher Entry counts suggest more dynamic scanning; higher SGE values indicate a more dispersed gaze distribution, and increased GTE values reflect a more complex sequential scanning pattern. On the contrary, during urgent tasks that require rapid identification of a specific target, such as detecting the alert and figuring out the next action in the TW2, the better group tends to demonstrate significantly lower Entry counts and reduced entropy values compared to the other group. Lower Entry counts suggest less visual scanning; higher SGE values indicate a more focused gaze distribution, and increased GTE values reflect a more assured and predictable sequential scanning pattern.

11.1.1.2 The driving study (Comparison A.2)

Similar insights were found in the driving domain, where three different TWs were identified, as introduced in Chapter 4.3 — Pre-TOR (TW1), from TOR to hazard perceived (TW2), and from hazard perceived to the end (TW3). In TW1, drivers were responsible for monitoring the autonomous driving control system and remaining prepared for a possible TOR. In the TW2, their were required to resume control and detect any hazards that prompted the TOR. In the final TW, drivers were expected to manually operate the vehicle and ensure safe driving.

In Topic 1, drivers' eye movements were compared between shorter-TOT and longer-TOT groups, while in Topic 2, the same eye tracking features were examined between younger and older drivers. Similar eye movement behaviors were observed between shorter-TOT driver and younger driver groups, across three TWs. As discussed in 5.3.2.2, shorter-TOT drivers exhibited significantly fewer Entry counts in multiple AOIs throughout all three TWs. Regarding entropy, shorter-TOT drivers showed less GTE in the TW2.

Younger drivers exhibited a comparable pattern, with significantly fewer Entry counts in multiple AOIs across three TWs (see Chapter 7.3.2) and significantly lower GTE observed in the first two TWs (see Chapter 7.3.3).

In the driving domain, age is not directly related to takeover performance, so it is difficult to claim that younger drivers' eye movement is always desirable. However, the observed similarity between younger drivers and those with shorter TOT suggests that younger drivers may exhibit better performance to some extent. Meanwhile, we could conclude that lower Entry counts did appear with lower GTE at the same time, showing that the shorter-TOT and younger driver cohorts did demonstrate less scanning with assured eye movement pattern together.

11.1.2 Comparison between driving and piloting domains (Comparison B)

From two sections above, the piloting experiment scenarios demonstrated three distinct eye movement patterns across the defined TWs. In first TW, the pilot engaged in routine monitoring of the autopilot system. In TW2, their visual behavior shifted into a focused and assured patterns, reflecting the need to detect alerts and determine an appropriate response. In TW3, following the disengagement of the autopilot, their eye movements returned to a dynamic scanning pattern.

In the driving domain, the difference between three TWs are less pronounced. However, operators from both domains exhibited a consistent eye movement pattern in the second TW. In this critical period, during which operators were required to detect an issue and resume control quickly, operators in the better groups (e.g. shorter-TOT driver and pilot, expert pilots and younger drivers) consistently exhibited fewer Entry counts and lower GTE. Fewer entry counts indicate reduced visual scanning during this period, and lower GTE values reflect a more focused and predictable eye movement pattern. This may result from more efficient eye movement patterns, where these operators were able to locate targets more quickly and scan surrounding visual information using an optimized scanning sequences.

In the first and third TW, the difference between the two domains is possibly due to the distinct nature of their autonomous systems. In the driving experiment, drivers operated a level-3 autonomous vehicle and were expected to remain ready for a takeover request. In contrast, pilots in the aviation experiment relied on the autopilot system for landing without being informed when a system malfunction might occur. In last TW, disparity in task time duration further influenced gaze pattern. For drivers, TW3 typically

spanned only 10–20 seconds, reflecting the immediate period after control was resumed. For pilots, TW3 lasted approximately 1–2 minutes, providing more time to resume systematic monitoring, which led to more pronounced monitoring-related eye movement patterns. Similar eye movement patterns were observed in previous research, where pilots' attention allocation, gaze dispersion, and gaze sequence complexity significantly decreased when facing an emergency situation [10]. This consistency suggests that under time pressure or emergency conditions, operators may adopt specific visual strategies.

These findings suggest that better operators are capable of adapting their gaze strategies based on task demands, shifting flexibly between dynamic scanning and concentrated distribution as needed. This adaptability highlights the importance of considering task context and time period when examining operators' eye movement. Evaluation metrics in previous studies often rely on eye movement measures without accounting for the temporal windows during which operators may pursue distinctly different goals. Incorporating task-specific gaze behavior into assessments could lead to more accurate evaluations and more effective, targeted training interventions

11.2 Limitations and future work

For experimental studies, results are often constrained by the conditions and resources available at the time of data collection. Several limitations observed in this work are listed below. I believe the findings and conclusions of this thesis could be further improved by address the following future directions, each corresponding to a specific limitation:

1. **Eye tracking**

   In the current thesis, both driving and piloting studies used head-mounted eye tracking glasses to record operators' eye movements, and features extracted from these recordings provided all subsequent analyses. Therefore the eye tracking data constitute the core part of the current research. Nonetheless, these data are subject to the following limitations:

   (a) **Eye tracking data quality issues** may have affected the accuracy of the findings. Occasional data loss occurred due to blinking, occlusion, unexpected movement of the eye tracker, or drift in the Inertial Measurement Unit (IMU). In addition, data quality varied across participants, often influenced by individual differences such glasses, dry eyes, or droopy eyelids, potentially leading to inconsistency in the recorded eye movement data.

(b) **Inaccurate AOI mapping** represents another limitation. The study relied on the detection of QR codes to locate AOI; however, this process was occasionally unreliable. QR codes were not consistently detected due to occlusion or extreme viewing angles, particularly when participants sat too close to the dashboard or when QR codes fell outside the field of view. These issues may have resulted in mis-classification of gaze points in terms of which AOIs they land in.

(c) **Differences in eye tracking hardware** between the driving and piloting studies introduced inconsistency in the recorded eye tracking data. The two eye trackers differed in fixation and saccade detection algorithms, sampling rates, and calibration procedures. These hardware-related differences may have led to subtle but meaningful inconsistencies, limiting direct comparisons across domains and participants.

(d) **Using eye tracking to measure SA** remains context-dependent. Although the study employed a validated set of eye tracking features as proxies for SA, their effectiveness varies across experimental contexts. Prior research supports the link between gaze behavior and SA, but findings are often influenced by the specific SA measurement method (e.g., SAGAT versus subjective ratings) and experimental design (e.g., task complexity, interface layout). Therefore, while eye tracking offers a promising non-intrusive approach to assessing SA, its generalizability should be interpreted with caution.

To address this issue, we continuously monitored the eye tracking video recordings during data collection and verified the gaze accuracy, measured by the Mean Angular Error, before and after each session. If the accuracy exceeded acceptable thresholds, we repeated the calibration procedure with the participant. In the future work, data accuracy could be further improved by using a more stable eye tracking system or applying filtering algorithms during analysis.

2. **Simulator setups**

(a) **Driving simulator fidelity** could be improved. In the current setup, only a single monitor was used to present the driving scenario. Some participants, especially older drivers, performed involuntary shoulder checks, which suggests that the limited visual field may have constrained their natural behaviors. Future studies could benefit from a more immersive setup, such as using three surrounding screens or a high-fidelity driving simulator (e.g., the VI-grade driving simulator). These enhancements may yield more valid results.

(b) **Autopilot functionality** was only limited to ILS in the piloting experiment, representing just one basic function of the AFCS. This limitation is partly due to the capabilities of the simulator used. However, modern autopilot systems comprise many additional functions (e.g., altitude hold, route management, automated go-around), which are more commonly used in modern aviation compared to the ILS landing (many airplanes use GPS landing now). Future studies could explore a wider variety of automated procedures to understand how different autopilot functionalities influence SA and pilot behavior.

3. **SAGAT questions**

   (a) **Timing of SAGAT questions** was set after landing in the piloting studies. While this approach allowed for a more comprehensive observation of post-takeover eye movement, it only allowed researchers to ask SAGAT questions at the very end of each trial. This represents a methodological trade-off. In future work, different results may be observed if the scenario is terminated earlier during the landing phase to reflect more valid SA of pilots.

   (b) **SAGAT Questions Design** is challenging in both domains. While SAGAT provides an objective assessment of operators' SA, designing high-quality, valid questions remains a challenge. Moreover, there is a lack of transparency in the literature regarding how SAGAT questions are developed and validated. Future research should focus on improving question design and reporting practices to ensure a more comprehensive coverage of operators' SA.

4. **Modeling**

   (a) **Relatively small sample sizes** may constrain the performance of the machine learning models developed in this study, leading to overfitting and reduced generalizability. In addition, the binary group segmentation (e.g., better versus worse SA) was based on cutoffs that may be considered arbitrary. In the future, a larger sample would allow more robust data-driven clustering and potentially reveal more robust group distributions.

   (b) **Generalized Linear Models (GLM)** were initially considered for modeling, but they were ultimately excluded due to unsatisfactory model performance in preliminary tests. For future work, increasing the sample size and incorporating more valid SAGAT questions may help strengthen the relationship between eye tracking data and the metrics used to measure SA and takeover performance.

11.3 Applications and Implications

In the current study, three research topics surrounding the eye movement of operators during the takeover process were investigated.

Although takeover scenarios have been widely studied in the driving domain, the transition from autopilot to manual flying has been much less explored. The piloting experiment in this thesis introduced the use of an ILS to simulate a takeover scenario in aviation. The results confirmed that pilots' eye movement patterns exhibited significant difference between different groups, as well as across the three event-based TWs defined in this study. Future research can build upon this work by conducting more piloting experiments under diverse automation failure scenarios to further examine how pilots reallocate attention during critical transitions.

A comprehensive set of eye-tracking features was developed and extracted within the three TWs. Significant group-level differences were observed in multiple AOI-based features across these TWs, underscoring the value of using temporally segmented eye tracking features. These results imply that analyzing AOI features repeatedly across TW is an effective approach for understanding how operators gather and process information during dynamic tasks. Future research could extend this approach to investigate eye movement patterns not only within individual TW but also in terms of how operators shift their strategies along time, enabling deeper insight into temporal information-gathering behaviors. These methodological contributions of this thesis provide a foundation for future research that seeks to evaluate operator SA and takeover performance more systematically.

Together, the results across Chapters 5–10 and the cross-domain synthesis presented in this chapter contribute several key implications. First, the recurring patterns observed in features such as Entry Count and GTE between TW1 and TW3 suggest that pilots who achieve shorter takeover times adopt consistent information-gathering strategies across monitoring and manual control phases. Similarly, this situation of higher Entry counts in multiple AOIs was also found for expert pilot in TW1, further confirming that active scanning is an indicator of a preferred eye movement pattern during monitoring phrase. These findings contribute to the theoretical understanding of pilots' visual attention management, offering a potential framework for defining "preferred" or optimal eye movement strategies.

Beyond theoretical implications, the findings of this thesis have practical applications for both training and system design. For training, the identified preferred eye movement patterns can inform the development of instructional programs that teach operators to adopt more effective visual scanning strategies, potentially improving their SA and takeover

performance. For system design, the insights into temporal attention allocation can guide the creation of adaptive TWs that better support operators during critical transitions, for example by providing earlier or context-aware alerts that match the operator's current visual state. Together, these applications may contribute to safer and more efficient human–automation interaction in both driving and piloting contexts.

References

[1] Measuring driver situation awareness using region-of-interest prediction and eye tracking | IEEE conference publication | IEEE xplore.

[2] Federal Aviation Administration. Automation. In *Risk Management Handbook*, pages 7–1–7–4. Washington, DC: Government Printing Office.

[3] Ingrid B. Anderson, Karin M. Bauer, Douglas W. Harwood, and Kay Fitzpatrick. Relationship to safety of geometric design consistency measures for rural two-lane highways. 1658(1):43–51. Publisher: SAGE Publications Inc.

[4] K Anstey, J Wood, S Lord, and J Walker. Cognitive, sensory and physical factors enabling driving safety in older adults. 25(1):45–65.

[5] Peren Arin, Maria Minniti, Samuele Murtinu, and Nicola Spagnolo. Inflection points, kinks, and jumps: A statistical approach to detecting nonlinearities. *Organizational Research Methods*, 25(4):786–814, 2022.

[6] Hameed Aswad Mohammed, David S. Hurwitz, and Kristen L. Macuga. Drivers' visual attention during the onset of the circular yellow indication at high-speed signalized intersections. 21(4):259–264. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15389588.2020.1741559.

[7] Autonomous Vehicle International. Mercedes-benz approved for 95km/h level 3 autonomous driving in germany, 2024. Accessed: April 29, 2025.

[8] Naila Ayala, Suzanne Kearns, Elizabeth Irving, Shi Cao, and Ewa Niechwiej-Szwedo. The effects of a dual task on gaze behavior examined during a simulated flight in low-time pilots. *Frontiers in Psychology*, 15:1439401, 2024.

[9] Naila Ayala, Suzanne Kearns, Elizabeth Irving, Shi Cao, and Ewa Niechwiej-Szwedo. Investigating the role of flight phase and task difficulty on low-time pilot performance, gaze dynamics and subjective situation awareness during simulated flight. *Journal of Eye Movement Research*, 17(1):10–16910, 2024.

[10] Naila Ayala, Suzanne Kearns, Elizabeth Irving, Shi Cao, and Ewa Niechwiej-Szwedo. Eye tracking during a complex aviation task for insights into information processing. *Journal of Visualized Experiments (JoVE)*, (218):e66886, 2025.

[11] Naila Ayala, Diako Mardanbegi, Abdullah Zafar, Ewa Niechwiej-Szwedo, Shi Cao, Suzanne Kearns, Elizabeth Irving, and Andrew T Duchowski. Does fiducial marker visibility impact task performance and information processing in novice and low-time pilots? *Computers & Graphics*, 119:103889, 2024.

[12] Naila Ayala, Abdullah Zafar, Suzanne Kearns, Elizabeth Irving, Shi Cao, and Ewa Niechwiej-Szwedo. The effects of task difficulty on gaze behaviour during landing with visual flight rules in low-time pilots. *Journal of eye movement research*, 16(1):10–16910, 2023.

[13] Ali Bahrami. Us department of transportation. 2000.

[14] Jiming Bai, Xu Sun, Shi Cao, Qingfeng Wang, and Jiang Wu. Exploring the timing of disengagement from nondriving related tasks in scheduled takeovers with pre-alerts: an analysis of takeover-related measures. *Human factors*, 66(12):2669–2690, 2024.

[15] Victoria A Banks, Katherine L Plant, and Neville A Stanton. Driving aviation forward; contrasting driving automation and aviation automation. *Theoretical issues in ergonomics science*, 20(3):250–264, 2019.

[16] Roxanne S Beltran and Corey E Tarwater. Overcoming the pitfalls of categorizing continuous variables in ecology, evolution and behaviour. *Proceedings B*, 291(2032):20241640, 2024.

[17] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[18] Dennis B. Beringer and Howard C. Harris, Jr. Automation in general aviation: Two studies of pilot responses to autopilot malfunctions. 9(2):155–174.

[19] Horatiu Boeriu. Improved highway assistant: One of the best features in the bmw ix m60, October 2023. Accessed: 2025-04-29.

[20] Stephanie Brams, Rafaël F. Rejtman, Oron Levin, Gal Ziv, Ignace TC Hooge, and Werner F. Helsen. Eye-tracking in aviation: a new method for detecting learned visual scan patterns of cockpit instrument in simulated flight. page 9 p. Artwork Size: 9 p. Medium: application/pdf Publisher: ETH Zurich.

[21] Stefan Brandenburg and Lewis Chuang. Take-over requests during highly automated driving: How should they be presented and under what conditions? 66:214–225.

[22] Shi Cao, Siby Samuel, Yovela Murzello, Wen Ding, Xuemei Zhang, and Jianwei Niu. Hazard perception in driving: A systematic literature review. *Transportation research record*, 2676(12):666–690, 2022.

[23] David B Carr and Prateek Grover. The role of eye tracking technology in assessing older driver safety. *Geriatrics*, 5(2):36, 2020.

[24] Benjamin T. Carter and Steven G. Luke. Best practices in eye tracking research. 155:49–62.

[25] Gabriela Ciuperca and Valerie Girardin. On the estimation of the entropy rate of finite markov chains. pages 1109–1117.

[26] Hallie Clark and Jing Feng. Age differences in the takeover of vehicle control and engagement in non-driving-related activities in simulated driving with conditional automation. 106:468–479.

[27] John D. Corrigan and Nancy S. Hinkeldey. Relationships between parts a and b of the trail making test. 43(4):402–409. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-4679%28198707%2943%3A4%3C402%3A%3AAID-JCLP2270430411%3E3.0.CO%3B2-E.

[28] Nelson Cowan. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. 104(2):163–191. Place: US Publisher: American Psychological Association.

[29] J. C. F. de Winter, Y. B. Eisma, C. D. D. Cabrall, P. A. Hancock, and N. A. Stanton. Situation awareness based on eye movements in relation to the task environment. 21(1):99–111.

[30] Joost C. F. de Winter, Riender Happee, Marieke H. Martens, and Neville A. Stanton. Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. 27:196–217.

[31] Frederic Dehais, Vsevolod Peysakhovich, Sébastien Scannella, Jennifer Fongue, and Thibault Gateau. "automation surprise" in aviation: Real-time solutions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2525–2534. ACM.

[32] Frédéric Dehais, Stéphane Juaneda, and Vsevolod Peysakhovich. Monitoring eye movements in real flight conditions for flight training purpose. pages 54–60.

[33] Chao Deng, Shi Cao, Chaozhong Wu, and Nengchao Lyu. Modeling driver take-over reaction time and emergency response time using an integrated cognitive architecture. *Transportation research record*, 2673(12):380–390, 2019.

[34] Hanwen Deng, Guoliang Xiang, Jiandong Pan, Xianhui Wu, Chaojie Fan, Kui Wang, and Yong Peng. How to design driver takeover request in real-world scenarios: A systematic review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 104:411–432, 2024.

[35] J Deniel, D Schnebelen, E Reynaud, MC Ouimet, and J Navarro. Driver's gaze behaviour before, during and after take-over manoeuvres: Influence of agentivity associated with different automation solutions. *Transportation research part F: traffic psychology and behaviour*, 100:181–196, 2024.

[36] Charlotte Desmet and Kevin Diependaele. An eye-tracking study on the road examining the effects of handsfree phoning on visual attention. 60:549–559.

[37] W Ding, Y Murzello, S Samuel, and S Cao. Examining the takeover process for senior drivers in conditionally automated vehicles. *Advances in Transportation Studies*, 2024.

[38] Wen Ding, Yovela Murzello, Shi Cao, and Siby Sameul. Where to gaze during takeover: Eye gaze strategy analysis of different situation awareness and hazard perception levels. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, page 21695067231193650. SAGE Publications Sage CA: Los Angeles, CA, 2023.

[39] Wen Ding, Yovela Murzello, Shi Cao, and Siby Samuel. Exploring the relationship between drivers' stationary gaze entropy and situation awareness in a level-3 automation driving simulation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 879–884. SAGE Publications Sage CA: Los Angeles, CA, 2024.

[40] Wen Ding, Yovela Murzello, Siby Samuel, and Shi Cao. Examining driver situation awareness in the takeover process of conditionally automated driving with the effect of age. *IEEE Access*, 12:96169–96178, 2024.

[41] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16. PMLR. ISSN: 2640-3498.

[42] Tania Dukic and Thomas Broberg. Older drivers' visual search behaviour at intersections. *Transportation research part F: traffic psychology and behaviour*, 15(4):462–470, 2012.

[43] Francis T. Durso, Carla A. Hackworth, Todd R. Truitt, Jerry Crutchfield, Danko Nikolic, and Carol A. Manning. Situation awareness as a predictor of performance for en route air traffic controllers. 6(1):1–20. Publisher: American Institute of Aeronautics and Astronautics, Inc.

[44] Mica Endsley. A taxonomy of situation awareness errors, human factors in aviation operations;. 3:287–292.

[45] Mica R. Endsley. Design and evaluation for situation awareness enhancement. 32(2):97–101. Publisher: SAGE Publications.

[46] Mica R Endsley. Situation awareness in aviation systems. (11):257–276.

[47] Mica R. Endsley. Toward a theory of situation awareness in dynamic systems. 37(1):32–64. Publisher: SAGE Publications Inc.

[48] Mica R. Endsley and Daniel J. Garland. *Situation Awareness Analysis and Measurement*. CRC Press. Google-Books-ID: WrJGDsjJakcC.

[49] Mica R. Endsley and Esin O. Kiris. The out-of-the-loop performance problem and level of control in automation. 37(2):381–394. Publisher: SAGE Publications Inc.

[50] Mica R. Endsley, Stephen J. Selcon, Thomas D. Hardiman, and Darryl G. Croft. A comparative analysis of sagat and sart for evaluations of situation awareness. 42(1):82–86. Publisher: SAGE Publications Inc.

[51] M.R. Endsley. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, pages 789–795 vol.3.

[52] Arye R. Ephrath and Renwick E. Curry. Detection by pilots of system failures during instrument landings. 7(12):841–848. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics.

[53] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. 41(4):1149–1160.

[54] Martin L. Fracker. Attention allocation in situation awareness. 33(20):1396–1400. Publisher: SAGE Publications.

[55] Adam J Garfein and A Regula Herzog. Robust aging among the young-old, old-old, and oldest-old. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 50(2):S77–S87, 1995.

[56] Christian Gold, Daniel Damböck, Lutz Lorenz, and Klaus Bengler. "take over!" how long does it take to get the driver back into the loop? 57(1):1938–1942. Publisher: SAGE Publications Inc.

[57] J. Golding. Motion sickness susceptibility questionnaire short-form (MSSQ-short). 47:507–516.

[58] Zbigniew Gomolka, Damian Kordos, and Ewa Zeslawska. The application of flexible areas of interest to pilot mobile eye tracking. 20(4):986. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[59] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1994.

[60] Jonas Gouraud, Arnaud Delorme, and Bruno Berberian. Autopilot, mind wandering, and the out of the loop performance problem. *Frontiers in neuroscience*, 11:541, 2017.

[61] Leo Gugerty. *Handbook for driving simulation in engineering, medicine and psychology*.

[62] Raija Hämäläinen, Bram De Wever, Katriina Sipiläinen, Ville Heilala, Arto Helovuo, Sami Lehesvuori, Miitta Järvinen, Jouni Helske, and Tommi Kärkkäinen. Using eye tracking to support professional learning in vision-intensive professions: a case of aviation pilots. *Education and Information Technologies*, 29(18):24803–24833, 2024.

[63] Sogand Hasanzadeh, Behzad Esmaeili, and Michael D Dodd. Examining the relationship between construction workers' visual attention and situation awareness under fall and tripping hazard conditions: Using mobile eye tracking. *Journal of construction engineering and management*, 144(7):04018060, 2018.

[64] Andreas Haslbeck and Hans-Juergen Hoermann. Flying the needles: flight deck automation erodes fine-motor flying skills among airline pilots. *Human factors*, 58(4):533–545, 2016.

[65] Robin Haynes, Iain R. Lake, Simon Kingham, Clive E. Sabel, Jamie Pearce, and Ross Barnett. The influence of road curvature on fatal crashes in new zealand. 40(3):843–850.

[66] Indra Herdiana, M Alfin Kamal, Mutia Nur Estri, et al. A more precise elbow method for optimum k-means clustering. *arXiv preprint arXiv:2502.00851*, 2025.

[67] Sebastian Hergeth, Lutz Lorenz, and Josef F. Krems. Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. 59(3):457–470. Publisher: SAGE Publications Inc.

[68] DAVID N. HOGG, KNUT FOLLES, FRODE STRAND-VOLDEN, and BELÉN TORRALBA. Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms. 38(11):2394–2413. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00140139508925275.

[69] M. S. Horswill, S. A. Marrington, C. M. McCullough, J. Wood, N. A. Pachana, J. McWilliam, and M. K. Raikos. The hazard perception ability of older drivers. 63(4):P212–P218.

[70] Chao Huang, Bo Yang, and Kimihiko Nakano. Impact of duration of monitoring before takeover request on takeover time with insights into eye tracking data. *Accident Analysis & Prevention*, 185:107018, 2023.

[71] Gaojian Huang and Brandon Pitts. Driver-vehicle interaction: The effects of physical exercise and takeover request modality on automated vehicle takeover performance

between younger and older drivers. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–4.

[72] Gaojian Huang and Brandon J. Pitts. Takeover requests for automated driving: The effects of signal direction, lead time, and modality on takeover performance. 165:106534.

[73] Gaojian Huang and Brandon J. Pitts. To inform or to instruct? an evaluation of meaningful vibrotactile patterns to support automated vehicle takeover performance. 53(4):678–687. Conference Name: IEEE Transactions on Human-Machine Systems.

[74] Hyundai Motor Group. Hyundai motor group executive chair advocates 'trust by taking on challenges and making a new leap through change' in new year's message, 2023. Accessed: 2025-05-26.

[75] SAE International. J3016: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles.

[76] H.R. Jex, J.D. McDonnell, and A.V. Phatak. A "critical" tracking task for manual control research. HFE-7(4):138–145. Conference Name: IEEE Transactions on Human Factors in Electronics.

[77] Lesong Jia, Chenglue Huang, and Na Du. Drivers' situational awareness of surrounding vehicles during takeovers: Evidence from a driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 106:340–355, 2024.

[78] Yunyi Jia, Longxiang Guo, and Xin Wang. 4 - real-time control systems. In Lipika Deka and Mashrur Chowdhury, editors, *Transportation Cyber-Physical Systems*, pages 81–113. Elsevier.

[79] Kaori Kawabata, Yuya Nakajima, Kazuki Fujita, Mamiko Sato, Koji Hayashi, and Yasutaka Kobayashi. Pilot study on gaze characteristics of older drivers while watching driving movies. *Geriatrics*, 9(5):132, 2024.

[80] Young Woo Kim and Sol Hee Yoon. Understanding drivers' situation awareness in highly automated driving using sagat, sart, and eye-tracking data. *Transportation Research Part F: Traffic Psychology and Behaviour*, 109:1437–1450, 2025.

[81] Wayne K. Kirchner. Age differences in short-term retention of rapidly changing information. 55(4):352–358. Place: US Publisher: American Psychological Association.

[82] Donald Knuth. *The TEXbook*. Addison-Wesley, Reading, Massachusetts, 1986.

[83] Robert Krueger, Steffen Koch, and Thomas Ertl. Saccadelenses: interactive exploratory filtering of eye tracking trajectories. In *2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)*, pages 31–34.

[84] A Kuipers, A Kappers, CR Vanholten, JHW Vanbergen, and WJ Oosterveld. Spatial disorientation incidents in the rnlaf f 16 and f 5 aircraft and suggestions for prevention. *AGARD, Situational Awareness in Aerospace Operations 16 p(SEE N 90-28972 23-53)*, 1990.

[85] Masafumi Kunishige, Hiroshi Fukuda, Tadayuki Iida, Nami Kawabata, Chinami Ishizuki, and Hideki MIyaguchi. Spatial navigation ability and gaze switching in older drivers: A driving simulator study. *Hong Kong journal of occupational therapy*, 32(1):22–31, 2019.

[86] Thomas Köhn, Matthias Gottlieb, Michael Schermann, and Helmut Krcmar. Improving take-over quality in automated driving by interrupting non-driving tasks. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 510–517. ACM.

[87] Moritz Körber, Thomas Weißgerber, Luis Kalb, Christoph Blaschke, and Mehdi Farid. Prediction of take-over time in highly automated driving by two psychometric tests. 82(193):195–201. Publisher: 2006, Revista DYNA.

[88] Keith Laing and David Welch. Gm's cruise gets more oversight after robotaxi dragging incident. *BNN Bloomberg*.

[89] Leslie Lamport. *LATEX — A Document Preparation System*. Addison-Wesley, Reading, Massachusetts, second edition, 1994.

[90] Olivier Lefrançois, Nadine Matton, and Mickaël Causse. Improving airline pilots' visual scanning and manual flight performance through training on skilled eye gaze strategies. 7(4):70.

[91] Shuo Li, Phil Blythe, Weihong Guo, and Anil Namdeo. Investigation of older driver's takeover performance in highly automated vehicles in adverse weather conditions. 12(9):1157–1165. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/iet-its.2018.0104.

[92] Wen-Chin Li, John J. H. Lin, Graham Braithwaite, and Matt Greaves. The development of eye tracking in aviation (ETA) technique to investigate pilot's cognitive processes of attention and decision-making. European Association for Aviation Psychology. Accepted: 2018-03-13T15:47:01Z.

[93] Wen-Chin Li, Jingyi Zhang, Samuel Court, Peter Kearney, and Graham Braithwaite. The influence of augmented reality interaction design on pilot's perceived workload and situation awareness. 92:103382.

[94] Nade Liang, Jing Yang, Denny Yu, Kwaku O. Prakah-Asante, Reates Curry, Mike Blommer, Radhakrishnan Swaminathan, and Brandon J. Pitts. Using eye-tracking to investigate the effects of pre-takeover visual engagement on situation awareness during automated driving. 157:106143.

[95] Robert E. Llaneras, Jeremy Salinger, and Charles A. Green. Human factors issues associated with limited ability autonomous driving systems: Drivers' allocation of visual attention to the forward roadway. 7(2013). Number: 2013 Publisher: University of Iowa.

[96] Shayne Loft, Daniel B. Morrell, and Samuel Huf. Using the situation present assessment method to measure situation awareness in simulated submarine track management. 2(1):33–48. Publisher: Inderscience Publishers Ltd.

[97] Christophe Lounis, Vsevolod Peysakhovich, and Mickaël Causse. Visual scanning strategies in the cockpit are modulated by pilots' expertise: A flight simulator study. 16(2):e0247061. Publisher: Public Library of Science.

[98] Yanyu Lu, Yiyuan Zheng, Zhen Wang, and Shan Fu. Pilots' visual scanning behaviors during an instrument landing system approach. *Aerospace Medicine and Human Performance*, 91(6):511–517, 2020.

[99] Masha Maltz and David Shinar. Eye movements of younger and older drivers. *Human factors*, 41(1):15–25, 1999.

[100] Daniel Martinez-Marquez, Sravan Pingali, Kriengsak Panuwatwanich, Rodney A Stewart, and Sherif Mohamed. Application of eye tracking technology in aviation, maritime, and construction industries: A systematic review. *Sensors*, 21(13):4289, 2021.

[101] Stefan Mattes and Anders Hallén. Surrogate distraction measurement techniques: The lane change test. pages 107–121. Publisher: CRC Press Boca Raton, Fla.

[102] Rollin Mccraty and Fred Shaffer. Heart rate variability: New perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. 4(1):46–61. Publisher: SAGE Publications Inc.

[103] Donald McLean. Automatic flight control systems. 36(6):172–175. Publisher: SAGE Publications Ltd.

[104] Ranjana K. Mehta, S. Camille Peres, Ashley E. Shortz, Wimberly Hoyle, Melissa Lee, Gurtej Saini, Hong-Chih Chan, and Mitchell W. Pryor. Operator situation awareness and physiological states during offshore well control scenarios. 55:332–337.

[105] Fanxing Meng and Charles Spence. Tactile warning signals for in-vehicle systems. 75:333–346.

[106] Natasha Merat, Bobbie Seppelt, Tyron Louw, Johan Engström, John D Lee, Emma Johansson, Charles A Green, Satoshi Katazaki, Chris Monk, Makoto Itoh, et al. The "out-of-the-loop" concept in automated driving: proposed definition, measures and implications. *Cognition, Technology & Work*, 21(1):87–98, 2019.

[107] Mercedes-Benz Group. Mercedes-Benz DRIVE PILOT further expands U.S. availability: California certification, 2023. Accessed: 2025-04-29.

[108] Lisa J Molnar, Anuj K Pradhan, David W Eby, Lindsay H Ryan, Renee M St Louis, Jennifer Zakrajsek, Brittany Ross, Brian Lin, Chen Liang, Bethany Zalewski, et al. Age-related differences in driver behavior associated with automated vehicles and the transfer of control between automated and manual control: A simulator evaluation. Technical report, University of Michigan, Ann Arbor, Transportation Research Institute, 2017.

[109] Walter Morales-Alvarez, Oscar Sipele, Régis Léberon, Hadj Hamma Tadjine, and Cristina Olaverri-Monreal. Automated driving: A literature review of the take over request in conditional automation. 9(12):2087. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[110] Endsley MR. A systematic review and meta-analysis of direct objective measures of situation awareness: A comparison of sagat and spam. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 63(1):124–150, 2021.

[111] Céline M. Muehlethaler and Chiara P. Knecht. Situation awareness training for general aviation pilots using eye tracking. 49(19):66–71.

[112] Muhammad Nanang Prayudyanto, Arief Goeritno, Safaruddin Ikhsan, and Fadhila Taqwa. Designing a model of the early warning system on the road curvature to prevent the traffic accidents.

[113] Frederik Naujoks, Dennis Befelein, Katharina Wiedemann, and Alexandra Neukum. A review of non-driving-related tasks used in studies on automated driving. In Neville A Stanton, editor, *Advances in Human Aspects of Transportation*, Advances in Intelligent Systems and Computing, pages 525–537. Springer International Publishing.

[114] Frederik Naujoks, Christoph Mai, and Alexandra Neukum. The effect of urgency of take-over requests during highly automated driving under distraction conditions. volume 7.

[115] Thanh Nguyen, Chee Peng Lim, Ngoc Duy Nguyen, Lee Gordon-Brown, and Saeid Nahavandi. A review of situation awareness assessment approaches in aviation environments. 13(3):3590–3603. Conference Name: IEEE Systems Journal.

[116] Sina Nordhoff, John D. Lee, Simeon C. Calvert, Siri Berge, Marjan Hagenzieker, and Riender Happee. (Mis-)use of standard Autopilot and Full Self-Driving (FSD) Beta: Results from interviews with users of Tesla's FSD Beta. *Frontiers in Psychology*, Volume 14 - 2023, 2023.

[117] S. D. Norman. *Flightdeck Automation: Promises and Realities*. NASA. Google-Books-ID: 4M0hAQAAIAAJ.

[118] Ministry of Transport. Aircraft accident investigation report.

[119] Chester L Olson. Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 69(348):894–908, 1974.

[120] International Civil Aviation Organization. ICAO annex 10 volume 1, radio navigation aids.

[121] Erfan Pakdamanian, Shili Sheng, Sonia Baee, Seongkook Heo, Sarit Kraus, and Lu Feng. DeepTake: Prediction of driver takeover behavior using multimodal data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–14. Association for Computing Machinery.

[122] E. H. J. Pallet. *Automatic Flight Control*. Blackwell Science, 4 edition.

[123] Hengyan Pan, William Payre, Jinhua Xu, and Sjaan Koppel. Age-related differences in takeover performance: A comparative analysis of older and younger drivers in prolonged partially automated driving. *Traffic injury prevention*, 25(7):968–975, 2024.

[124] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. A model for types and levels of human interaction with automation. 30(3):286–297.

[125] Qijia Peng and Sunao Iwaki. Visual attention of young and older drivers in takeover tasks of highly automated driving. In *International conference on human-computer interaction*, pages 210–221. Springer, 2020.

[126] Sebastiaan Petermeijer, Pavlo Bazilinskyy, Klaus Bengler, and Joost de Winter. Take-over again: Investigating multimodal and directional TORs to get the driver back into the loop. 62:204–215.

[127] Jonas Radlmayr, Christian Gold, Lutz Lorenz, Mehdi Farid, and Klaus Bengler. How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving. 58(1):2063–2067. Publisher: SAGE Publications Inc.

[128] Keith Rayner. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. 62(8):1457–1506. Publisher: SAGE Publications.

[129] Keith Rayner. Eye movements in reading and information processing: 20 years of research. 124(3):372–422. Place: US Publisher: American Psychological Association.

[130] CIO-40 Records Management Division. Descent below visual glidepath and impact with seawall asiana airlines flight 214.

[131] Umair Rehman, Shi Cao, and Carolyn MacGregor. Using an integrated cognitive architecture to model the effect of environmental complexity on drivers' situation awareness. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 812–816. SAGE Publications Sage CA: Los Angeles, CA, 2019.

[132] Umair Rehman, Shi Cao, and Carolyn G MacGregor. Modelling level 1 situation awareness in driving: A cognitive architecture approach. *Transportation Research Part C: Emerging Technologies*, 165:104737, 2024.

[133] Maxime Reynal, Frank Rister, Sebastien Scannella, Christopher Wickens, and Frederic Hehais. Investigating pilot's decision making when facing an unstabilized approach: An eye-tracking study. pages 335–340.

[134] Mohammad Sadraey. *Automatic Flight Control Systems*. Morgan & Claypool Publishers. Google-Books-ID: 6qjRDwAAQBAJ.

[135] SAE International. J3016: Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems.

[136] Jason J. Saleem and Brian M. Kleiner. The effects of nighttime and deteriorating visual conditions on pilot performance, workload, and situation awareness in general aviation for both VFR and IFR approaches. 5(1):107–120. Publisher: Citeseer.

[137] L. Sanders and V. Fritch. Instrument landing systems. 21(5):435–454. Conference Name: IEEE Transactions on Communications.

[138] Nadine B. Sarter, Randall J. Mumaw, and Christopher D. Wickens. Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. 49(3):347–357.

[139] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.

[140] Fred Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. 5.

[141] C. E. Shannon. A mathematical theory of communication. 27(3):379–423. Conference Name: The Bell System Technical Journal.

[142] Khatereh Shariatmadari, Siby Samuel, Shi Cao, and Amandeep Singh. Comparison of two-level and three-level graded collision warning systems under distracted driving conditions. *IEEE Access*, 2025.

[143] Brook Shiferaw, Luke Downey, and David Crewther. A review of gaze entropy as a measure of visual scanning efficiency. *Neuroscience & Biobehavioral Reviews*, 96:353–366, 2019.

[144] Qian Sun, Jianhong Xia, Torbjörn Falkmer, and Hoe Lee. Investigating the spatial pattern of older drivers' eye fixation behaviour and associations with their visual capacity. *Journal of Eye Movement Research*, 9(6):30, 2016.

[145] Xiaomei Tan and Yiqi Zhang. The effects of takeover request lead time on drivers' situation awareness for manually exiting from freeways: A web-based study on level 3 automated vehicles. 168:106593.

[146] R. M. Taylor. Situational awareness rating technique (sart): The development of a tool for aircrew systems design. In *Situational Awareness*, pages 111–128. Routledge. Num Pages: 18.

[147] Julian F. Thayer, Fredrik Åhs, Mats Fredrikson, John J. Sollers, and Tor D. Wager. A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. 36(2):747–756.

[148] Louis Tijerina, Sam Johnston, E. Parmer, H. A. Pham, M. D. Winterbottom, Frank S. Barickman, Vehicle Research and Test Center (U.S.), and University of Colorado at Denver. Preliminary studies in haptic displays for rear-end collision avoidance system and adaptive cruise control system applications.

[149] Antonella Toffetti, Ellen S. Wilschut, Marieke H. Martens, Anna Schieben, Amon Rambaldini, Natasha Merat, and Frank Flemisch. CityMobil: Human factor issues regarding highly automated vehicles on eLane. 2110(1):1–8. Publisher: SAGE Publications Inc.

[150] United States. Department of Transportation and United States. Department of Defense. 2001 federal radionavigation systems.

[151] Prabitha Urwyler, Nicole Gruber, René M Müri, Michael Jäger, Rahel Bieri, Thomas Nyffeler, Urs P Mosimann, and Tobias Nef. Age-dependent visual exploration during simulated day-and night driving on a motorway: a cross-sectional study. *BMC geriatrics*, 15(1):18, 2015.

[152] Jared Van Dam, Steven J. Kass, and Lisa VanWormer. The effects of passive mobile phone interaction on situation awareness and driving performance. 12(8):1007–1024. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19439962.2018.1564947.

[153] Koen van de Merwe, Henk van Dijk, and Rolf Zon. Eye movements as an indicator of situation awareness in a flight simulator experiment. 22(1):78–95. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10508414.2012.635129.

[154] Arie P. van den Beukel and Mascha C. van der Voort. The influence of time-criticality on situation awareness when retrieving human control after automated driving. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 2000–2005. ISSN: 2153-0017.

[155] Samuel J Vine, Liis Uiga, Aureliu Lavric, Lee J Moore, Krasimira Tsaneva-Atanasova, and Mark R Wilson. Individual reactions to stress predict performance during a critical aviation incident. *Anxiety, Stress, & Coping*, 28(4):467–477, 2015.

[156] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[157] Bernhard Wandtner, Gerald Schmidt, Nadja Schoemig, and Wilfried Kunde. Non-driving related tasks in highly automated driving - effects of task modalities and cognitive workload on take-over performance. In *AmE 2018 - Automotive meets Electronics; 9th GMM-Symposium*, pages 1–6.

[158] Ange Wang, Wenxin Shi, Dengbo He, and Hai Yang. Impact of multi-dimensional cognitive demands on takeover performance, physiological and eye-tracking measures in conditionally automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, 114:461–475, 2025.

[159] Qiuhong Wang, Haolin Chen, Jianguo Gong, Xiaohua Zhao, and Zhenlong Li. Studying driver's perception arousal and takeover performance in autonomous driving. *Sustainability*, 15(1):445, 2022.

[160] Waymo. Waymo - autonomous driving technology, 2025. Accessed: 2025-04-29.

[161] Bradley W. Weaver and Patricia R. DeLucia. A systematic review and meta-analysis of takeover performance during conditionally automated driving. 64(7):1227–1260. Publisher: SAGE Publications Inc.

[162] Bradley W Weaver and Patricia R DeLucia. A systematic review and meta-analysis of takeover performance during conditionally automated driving. *Human factors*, 64(7):1227–1260, 2022.

[163] Hengyang Wei, Damin Zhuang, Xiaoru Wanyan, and Qun Wang. An experimental analysis of situation awareness for cockpit display interface evaluation based on flight simulation. 26(4):884–889.

[164] Hannah White, David Large, Davide Salanitri, Gary Burnett, Anneka Lawson, and Elizabeth Box. Rebuilding drivers' situation awareness during take-over requests in level 3 automated cars.

[165] Christopher D. Wickens, Justin G. Hollands, Simon Banbury, and Raja Parasuraman. Automation and human performance. In *Engineering Psychology and Human Performance*, pages 377–404. Taylor & Francis Group, 4 edition.

[166] William Little and Ron McGivern. Chapter 13: Aging and the elderly, 2014. Accessed on 2025-06-20.

[167] Jing Yang, Nade Liang, Brandon J. Pitts, Kwaku O. Prakah-Asante, Reates Curry, Mike Blommer, Radhakrishnan Swaminathan, and Denny Yu. Multimodal Sensing and Computational Intelligence for Situation Awareness Classification in Autonomous Driving. *IEEE Transactions on Human-Machine Systems*, 53(2):270–281, April 2023. Conference Name: IEEE Transactions on Human-Machine Systems.

[168] Sol Hee Yoon and Yong Gu Ji. Non-driving-related tasks, workload, and takeover performance in highly automated driving contexts. 60:620–631.

[169] Sol Hee Yoon, Young Woo Kim, and Yong Gu Ji. The effects of takeover request modalities on highly automated car control transitions. 123:150–158.

[170] Neng Zhang, Mohammad Fard, John Laurence Davy, Sibashis Parida, and Stephen R. Robinson. Is driving experience all that matters? drivers' takeover performance in conditionally automated driving.

[171] Ting Zhang, Jing Yang, Nade Liang, Brandon J. Pitts, Kwaku Prakah-Asante, Reates Curry, Bradley Duerstock, Juan P. Wachs, and Denny Yu. Physiological measurements of situation awareness: A systematic review. 65(5):737–758.

[172] Feng Zhou, X. Jessie Yang, and Joost C. F. de Winter. Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. 23(3):2284–2295. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

APPENDICES

Appendix A

SAGAT questions

Below are all the SAGAT questions used in the driving experiment. In the actual procedure, these questions were reused across different scenarios and reordered to prevent participants from encountering the same questions consecutively.

Highway scenario:

- Did you see any cars in your mirror?
- What speed was your car going at before the session ended?
- What speed was your car going at when you took over?
- Did you notice the pedestrian trying to cross the road?
- Were there any cars around you?
- How many lanes were present?
- Was there a speed limit sign?
- What was the speed limit posted on the sign before the scenario ended?
- Could you briefly describe the billboard sign overhead?
- What input did you first provide when disabling Autopilot?

Urban scenario:

- What speed was your car going at when the session ended?
- Did you see a stand on the left?
- Did you notice any traffic signals?
- Did you see any cars in your left mirror?
- What cars were in the opposite lane?
- Was there a bus stop?
- What was the speed limit posted on the sign before the scenario ended?
- Were there any cars ahead?
- What was the road speed limit?
- Was there a vehicle behind you in your rear-view mirror?
- What input did you first provide when disabling Autopilot?

Appendix B

Demographic Questionnaires

This questionnaire is strictly confidential. Only a randomly generated participant ID number, assigned by the research administrator, will appear on this form. No information provided will be traceable to you personally. You may skip any question you do not feel comfortable answering.

Participant ID (2-digit): _____

Gender (select one):

- Male
- Female
- Prefer not to say
- Prefer to self-identify: _____

Age (years): _____

Current Driver's License (select one):

- G
- G2

Total number of years of driving: _____

Estimated total driving distance (kilometers): _____

Experience with Semi-Autonomous Vehicles:

This is a vehicle that is capable of driving itself safely with little or no human input under certain conditions. The driver has the opportunity to engage in non-driving related tasks during vehicle automation but must be available to take over when the system reaches its limit. For example, Tesla Autopilot or the Traffic Jam Pilot feature in Audi A8 that autonomously steers, accelerates, and brakes the car when enabled.

- Yes
- No

If "Yes", please specify duration of usage (in months): _____

Appendix C

Motion Sickness Susceptibility Questionnaire Short-form

1. Please State Your **Age** ..... Years. *2.* Please State Your Sex (tick box) **Male** [ ] **Female** [ ]

This questionnaire is designed to find out how susceptible to motion sickness you are, and what sorts of motion are most effective in causing that sickness. Sickness here means feeling queasy or nauseated or actually vomiting.

Your CHILDHOOD Experience Only (before 12 years of age), for each of the following types of transport or entertainment please indicate:

3. **As a CHILD (before age 12),** how often you **Felt Sick or Nauseated** (tick boxes):

	Not Applicable - Never Travelled	**Never Felt Sick**	**Rarely Felt Sick**	**Sometimes Felt Sick**	**Frequently Felt Sick**
Cars					
Buses or Coaches					
Trains					
Aircraft					
Small Boats					
Ships, e.g. Channel Ferries					
Swings in playgrounds					
Roundabouts in playgrounds					
Big Dippers, Funfair Rides					

t 0 1 2 3

Your Experience over the LAST 10 YEARS (approximately), for each of the following types of transport or entertainment please indicate:

4. **Over the LAST 10 YEARS,** how often you **Felt Sick or Nauseated** (tick boxes):

	Not Applicable - Never Travelled	**Never Felt Sick**	**Rarely Felt Sick**	**Sometimes Felt Sick**	**Frequently Felt Sick**
Cars					
Buses or Coaches					
Trains					
Aircraft					
Small Boats					
Ships, e.g. Channel Ferries					
Swings in playgrounds					
Roundabouts in playgrounds					
Big Dippers, Funfair Rides					

t 0 1 2 3

Appendix D

Air traffic control scripts

To increase the diversity of the communication, there are two types of ATC scripts used in the piloting experiment: airplane started from Toronto tower and airplane started from Waterloo tower. Both categories are composed of "Other dialogue" which is about other air airplane, and "Relevant dialogue" which is about ego airplane. As shown in Figure D.1, those two dialogue are allocated to specific timing in the landing trajectory.

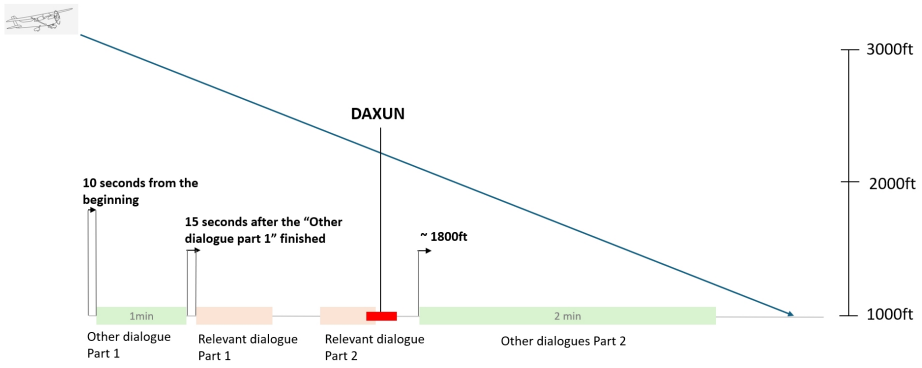


Figure D.1: ATC Dialogue Timing in Relation to the Landing Trajectory

ATC Scripts Type A: Airplane started from Toronto tower

Other Dialogue Part 1

C-GWEN: Toronto Centre, this is Golf-Whiskey-Echo-November, at 3000.
Toronto Centre: Whiskey-Echo-November, Toronto Centre, radar identified. Not below 2500, altimeter two-niner-niner-two.
C-GWEN: Not below 2500, altimeter two-niner-niner-two, Whiskey-Echo-November.

Toronto Centre: Romeo-Alpha-Foxtrot, Toronto Centre, contact City Tower on one-one-eight-decimal-two.
C-FRAF: Contact City Tower on one-one-eight-decimal-two, Romeo-Alpha-Foxtrot.

Toronto Centre: Whiskey-Echo-November, cleared to 5000.
C-GWEN: Cleared to 5000, Whiskey-Echo-November.

Relevant Dialogue Part 1

Toronto Centre: Whiskey-Alpha-Tango, contact Waterloo Tower on one-two-six-decimal-zero.
C-GWAT: Waterloo Tower on one-two-six-decimal-zero, Whiskey-Alpha-Tango.

C-GWAT: Waterloo Tower, Golf-Whiskey-Alpha-Tango, ILS Runway 26.
Waterloo Tower: Golf-Whiskey-Alpha-Tango, Waterloo Tower, Runway 26, wind 165 at 10, altimeter two-niner-niner-two, number two following a Diamond (A1) / Cirrus (A2) / Bonanza (A3) short final, report DAXUN.
C-GWAT: Whiskey-Alpha-Tango.

Relevant Dialogue Part 2

C-GWAT: Whiskey-Alpha-Tango, DAXUN final.
Waterloo Tower: Whiskey-Alpha-Tango, traffic is exiting at the end (A1) / at Charlie (A2) / at the main intersection (A3), wind 165 at 10, cleared to land Runway 26.
C-GWAT: Cleared to land Runway 26, Whiskey-Alpha-Tango.

Other Dialogue Part 2

C-GCAO: Waterloo Tower, this is Golf-Charlie-Alpha-Oscar with information Bravo.
Waterloo Tower: Charlie-Alpha-Oscar, Waterloo Tower, state your altitude and your intentions.
C-GCAO: Waterloo Tower, Charlie-Alpha-Oscar, 3000, inbound for circuits.
Waterloo Tower: Charlie-Alpha-Oscar, identified, cleared left downwind Runway 26, not below 2600.
C-GCAO: Cleared left downwind Runway 26, not below 2600, Charlie-Alpha-Oscar.

C-GSHI: Waterloo Tower, this is Golf-Sierra-Hotel-India with information Bravo.
Waterloo Tower: Sierra-Hotel-India, Waterloo Tower, state your altitude and your intentions.
C-GSHI: Waterloo Tower, Sierra-Hotel-India, 3000, inbound for full stop.
Waterloo Tower: Sierra-Hotel-India, identified, cleared for the Maryhill arrival, report over Maryhill.
C-GSHI: Cleared for the Maryhill arrival, will report over Maryhill, Sierra-Hotel-India.

ATC Scripts Type B: Airplane started from Waterloo tower

Other Dialogue Part 1

C-GWEN: Waterloo Tower, this is Golf-Whiskey-Echo-November, holding short 26, ready for departure.
Waterloo Tower: Whiskey-Echo-November, Waterloo Tower, wind 165 at 10, cleared takeoff Runway 26.
C-GWEN: Cleared takeoff Runway 26, Whiskey-Echo-November.
Waterloo Tower: Whiskey-Echo-November, contact Toronto Centre on one-two-eight-decimal-two-seven.
C-GWEN: Contact Toronto Centre on one-two-eight-decimal-two-seven, Whiskey-Echo-November.

Relevant Dialogue Part 1

Waterloo Tower: Golf-Whiskey-Alpha-Tango, Waterloo Tower, Runway 26, wind 165 at 10, altimeter two-niner-niner-two, continue number one, report DAXUN.
C-GWAT: Whiskey-Alpha-Tango.

Relevant Dialogue Part 2

C-GWAT: Tower, Whiskey-Alpha-Tango, DAXUN final.
Waterloo Tower: Whiskey-Alpha-Tango, vehicle crossing at Charlie (B1) / at the main intersection (B2) / at Echo (B3), wind 165 at 10, cleared to land Runway 26.
C-GWAT: Cleared to land Runway 26, Whiskey-Alpha-Tango.

Other Dialogue Part 2

C-GSHI: Waterloo Tower, this is Golf-Sierra-Hotel-India, holding short 26, ready for departure.
Waterloo Tower: Sierra-Hotel-India, Waterloo Tower, stand by, landing traffic.
C-GSHI: Standing by, Sierra-Hotel-India.

C-GCAO: Waterloo Tower, this is Golf-Charlie-Alpha-Oscar with information Bravo.
Waterloo Tower: Charlie-Alpha-Oscar, Waterloo Tower, state your altitude and your intentions.
C-GCAO: Waterloo Tower, Charlie-Alpha-Oscar, 3000, inbound for circuits.
Waterloo Tower: Charlie-Alpha-Oscar, identified, cleared left downwind Runway 26, not below 2600.
C-GCAO: Cleared left downwind Runway 26, not below 2600, Charlie-Alpha-Oscar.

Appendix E

SAGAT questions for piloting experiment

Below are all the SAGAT questions used in the piloting experiment. In the actual procedure, these questions were reused across different scenarios and reordered to prevent participants from encountering the same questions consecutively.

ATC-specific questions:

- Are you following an aircraft during landing? What is it?
- Which number were you in the landing sequence?
- What did the ATC instruct you to do at first?
- Where were you asked to exit?
- Where was the vehicle in front of you?
- What was happening on the radio before you landed?

Non-ATC-related questions:

- From beginning to roundouts, has your air speed gone out of 60-75 knots?
- How far down the runway did you land?
- What is the approach speed of your aircraft prior to landing?

- Did you land on the left or right of the centerline?
- At what altitude did the runway first become visible?
- At what distance did the runway first become visible?
- What is the altitude of the aircraft when malfunction happens?
- What is the airspeed of the aircraft when malfunction happens?
- What is the distance to the runway when the malfunction occurs?
- Why did the aircraft give an alert?
- Which happened to the autopilot function within this scenario?

Appendix F

Piloting demographic questionnaire

Participant ID: _____

If you have a pilot license, what is the type of license: _____

Do you hold your Instrument Rating:

- Yes
- No

What Other pilot rating do you have _____

How many flying hours you logged? _____

Will you be able to conduct ILS landing independently

- Yes
- No
- Maybe
- I don't know what ILS landing is