Designing Natural Interaction for Camera Viewpoint Control in Teleoperation

Dingyun Zhu

November 2011

A thesis submitted for the degree of Doctor of Philosophy of the Australian National University

Research School of Computer Science
College of Engineering and Computer Science
The Australian National University
To my family.

In memory of my grandfather.
Declaration

The work in this thesis is my own except where otherwise stated.

Dingyun Zhu
Dingyun Zhu
November 2011
Decision

The above information is for guidance purposes only.

[Signature]
[Date]
Acknowledgements

Standing on the shoulder of giants, I would like to express my sincere gratitude to all who have provided the support to complete this thesis. It is undoubted that without them this thesis would not have been what it is.

Firstly, I would like to gratefully thank my primary supervisor (and Chair of my supervisory panel): Professor Tom Gedeon who gave me the opportunity to choose my own path for my research and most importantly for guiding me along the way and educating me in the ways of academia and research. Tom was always very encouraging and helped me to look at problems from different perspectives. His philosophy of supervision has been "validated" by my research productivity. Working with him has been a wonderful experience throughout my entire PhD period. His boundless enthusiasm and open door have provided a creative setting for my work. Without his insightful knowledge, flexible guidance and endless patience, it would not have been possible for me to complete this research and I could not have expected a better supervisor for my PhD.

Secondly, I would like to thank Dr. Ken Taylor for being my CSIRO supervisor, who has spent a lot of effort helping with this thesis work: offering the chance to conduct the research on the remote mining project, contacting and interacting with both internal and external partners and collaborators, organising exhibitions and demonstrations to show off the prototype systems and providing so much valuable advice regarding teleoperation and robotics.

I would also like to express my appreciation to my university co-supervisor: Associate Professor Henry Gardner. With an expertise in Human Computer
Interaction, Henry has been consistently offering very useful advice particularly on interface design, conducting user evaluation and analysing data by correct statistical methods. His ideas and suggestions have led my thoughts through the entire research.

My special gratitude goes to Dr. Sumudu Mendis for being my advisor and more importantly as a mentor for my PhD. I would also like to thank many people from the Research School of Computer Science (RSCS) at the ANU: Nandita Sharma, Sukanya Manna, Baichen Chen, Torben Sko, Ben Swift, James Sheridan, Amir Hadad, Bodhi Philpot, Bagus Manuaba, Florian Nebout, Stephane Bersot, Sui Sin Khoo, Tan Vo, Viktor Gal, Akshay Asthana, Luke Nguyen-Hoan and Alvin Teh. It definitely has been a great time to work with them.

Matt Adcock and Chris Gunn from the Immersive Environments Team at the ICT Centre CSIRO helped me much on the software implementation and system integration for developing these prototypes, which saved me countless hours by sharing their professional programming and system construction expertise.

Thank you to the staff members from the CSIRO Minerals Down Under (MDU) National Research Flagship: Eleonora Widzyk-Capehart, Jock Cunningham and Anna Littleboy for their consistent support on this research.

Special thanks go to these great administrative people: Julie Arnold, Deanne Drummond and Chelsea Holton from ANU; Deirdre Tribe, Renate Hays, Sarah Savage, Daphne Bruce and Karen Barndt from CSIRO - without whom things would not have gone so smoothly.

I would also like to thank all volunteers who participated in the user studies. Without their participation and feedback, it would not have been possible to complete the evaluation of this research.

Last but not least, I am indebted to my parents who have always afforded me the freedom to follow my own path. I am aware that they have been missing me so much and would rather have me home, but they still encouraged me to follow my own research studies in Australia. My grandfather, Dazhang Zhu, who
had been remarkably fighting with cancers for the last two years of my PhD. He encouraged me to think free and pursue my own dreams. I dedicate this PhD to him for he would have been very proud.
Abstract

Teleoperation can be used to remove human beings or provide replaceable surrogates in hazardous or difficult working environments. Multi-tasking situations are common in teleoperation. For many teleoperation activities, operators usually have multiple devices or equipment to control via conventional hand operated interfaces, such as joysticks, switches, wheels, mice and keyboards. A typical example of such a situation is controlling a robot for either navigation or manipulation as the primary task and controlling a remote camera as the major visual feedback to obtain situational awareness from the remote environment, all by manual interfaces. Apart from switching hands frequently between different control interfaces, the working attention of the operator will also be distracted from the primary task under such circumstances, which results in increased workload and reducing performance.

This thesis is motivated by a real-world industrial setting in mining teleoperation: Remote rock breaking, in which operators have to face such a multi-tasking situation. Instead of using conventional control interfaces, we explore several possible solutions by integrating natural human interactive information to design and develop new remote camera viewpoint control models. Natural interaction based interfaces allow people to interact with technologies, applications and systems as they are used to when they interact with other people in everyday life through movements, gestures, expressions, discovering the world by looking around and manipulating physical objects. In this research, we particularly investigate the use of natural head movements, eye gaze, weight-based body gestures and foot
movements with prototype systems as alternative inputs for the camera viewpoint control in teleoperation.

We present a number of user studies which compare these camera viewpoint control models with conventional control interfaces. Several novel lab-based experimental models have been designed using the concept of a Functional Physical Model, which is used to simulate the real-world multi-tasking setting to deal with the difficulty of direct access to real industry settings. Both objective and subjective measures were used in all user studies conducted.

From the results, we demonstrate the remote viewpoint control models we developed are either comparable to or an improvement over conventional control interfaces used in the experiments. Particularly for the empirical user study comparing our Natural Interaction Model (combining eye gaze and head motion) with optimised manual control and autonomous tracking models, the results indicate the advantages of using our natural interaction model, while the manual control model performed the worst.

Existing issues, discovered problems and challenges of using natural interaction for remote camera viewpoint control are also discussed. Finally, suggestions are made for further research in the area of importing more natural interaction based interfaces for teleoperation.
List of Publications

The work described in this thesis is the original work of the author except where specific reference or acknowledgement is made to the work or contribution of others. Some of the material in this work has appeared previously in publications and presentations by the author or to which the author has contributed. In all cases, only the contribution made by the author has been included in this work unless specific reference to the contrary has been made.

Long Papers

   
   Video demonstration: [https://www.youtube.com/watch?v=ewRUlh48Aog](https://www.youtube.com/watch?v=ewRUlh48Aog)

   
   Video demonstration: [https://www.youtube.com/watch?v=QzbNc2dpT6I](https://www.youtube.com/watch?v=QzbNc2dpT6I)


**Short Papers**


Contents

Acknowledgements vii
Abstract xi
List of Publications xiii
List of Figures xxvi
List of Tables xxvii

1 Introduction 1
1.1 A Real-World Industrial Setting: Rock Breaking in Mining . . . . . . . 3
1.2 Thesis Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
1.3 Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
1.4 Organisation of the Dissertation . . . . . . . . . . . . . . . . . . . . . . . . 12

2 Background 17
2.1 Natural Interaction and User Interface . . . . . . . . . . . . . . . . . . . 17
2.1.1 What is Natural Interaction? . . . . . . . . . . . . . . . . . . . . . . . 18
2.1.2 History of Natural Interaction . . . . . . . . . . . . . . . . . . . . . . . 20
2.1.3 State of the Art in Natural Interaction . . . . . . . . . . . . . . . . . 21
2.1.3.1 Head and Face . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
2.1.3.2 Eye Gaze . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
2.1.3.3 Touch . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
2.1.3.4 Speech ........................................... 26
2.1.3.5 Gesture ......................................... 27

2.2 User Interface for Teleoperation ........................................... 29
  2.2.1 User Interface and Performance Issues ......................... 30
  2.2.2 User Interface for Mining Teleoperation ....................... 32
  2.2.3 Remote Camera Viewpoint Control .............................. 33
  2.2.4 Why Natural Interaction for Teleoperation? .................... 34

2.3 Summary .................................................. 36

3 Head Tracking Interface Design and Evaluation .......................... 37
  3.1 Related Work .............................................. 37
    3.1.1 Head Tracking Technologies and Systems .................... 37
      3.1.1.1 Sensor based Head Tracking .......................... 38
      3.1.1.2 Computer Vision based Head Tracking .................. 39
    3.1.2 Head Tracking for Viewpoint Control ....................... 40
  3.2 Our Design of Remote Camera Control ................................ 43
    3.2.1 Head Motion Control ................................... 43
    3.2.2 Head Flicking Control .................................. 44
    3.2.3 Zooming Control ........................................ 44
  3.3 Preliminary User Study ....................................... 46
    3.3.1 Apparatus and Implementation ................................ 46
    3.3.2 Participants ............................................ 46
    3.3.3 Experimental Design ..................................... 47
    3.3.4 Experimental Procedure .................................. 48
  3.4 Experimental Results ......................................... 49
    3.4.1 Objective Results ....................................... 49
    3.4.2 Subjective Results ..................................... 53
  3.5 Discussion .................................................. 53
  3.6 Learned Feedback and Improved Model .............................. 55
  3.7 Summary .................................................... 58
4  Eye Tracking Interface Design and Evaluation

4.1 Related Work ......................................................... 61
4.1.1 Eye Tracking Technologies and Systems ................. 61
4.1.2 Gaze Interaction and Applications .................. 64
4.1.3 Eye Tracking for Viewpoint Control ................ 66
4.2 Design of Gaze-Driven Remote Camera Control .......... 68
4.3 Prototype Implementation ........................................ 71
4.4 Evaluation ............................................................. 74
4.4.1 Experiment Setting Modeling ............................... 75
4.4.1.1 Functional Physical Modeling ....................... 76
4.4.1.2 Key Properties of Functional Physical Models ... 77
4.4.1.3 The Original Setting being Modeled .............. 78
4.4.2 Experimental Setup ............................................. 81
4.4.3 Participants ...................................................... 81
4.4.4 Experimental Design .......................................... 82
4.4.5 Experimental Procedure .................................... 83
4.5 Results ................................................................. 84
4.5.1 Objective Measure Results: User Performance on Goals
       and Kicks ...................................................... 84
4.5.2 Subjective Measure Results: User Preference and Feedback
       from Questionnaire and Interview ...................... 86
4.6 Discussion ............................................................ 90
4.6.1 Further Discussions and Observations on Results ...... 90
4.6.2 Existing Issues and Arguments ............................ 92
4.6.3 Summary of the Results .................................... 93
4.7 Summary ............................................................... 93

5  Further Explorations of Camera Viewpoint Control

5.1 Related Work ........................................................ 98
5.2 Camera Control Models Considered ............................ 99
### CONTENTS

5.2.1 Dual Manual Control - Conventional Interface ........ 100  
5.2.2 Natural Interaction - Eye Gaze + Head Motion .......... 102  
5.2.3 Autonomous Tracking .................................. 103  
5.3 User Study .................................................. 105  
  5.3.1 Apparatus and Implementation ....................... 105  
  5.3.2 Participants ........................................... 107  
  5.3.3 Experimental Design .................................. 108  
  5.3.4 Experimental Task and Procedure ................. 108  
5.4 Results ..................................................... 109  
  5.4.1 Objective Measures: Rocks and Nudges .......... 109  
  5.4.2 Subjective Measures: Questionnaire and Interview . 112  
5.5 Discussion .................................................. 116  
5.6 Summary ..................................................... 119  

6 Other Models and Interfaces ................................. 121  
6.1 More Models for Remote Camera Viewpoint Control .... 121  
  6.1.1 Related Work ........................................... 122  
  6.1.2 WiiFitChair: Body Gestures + Head Tracking .... 124  
  6.1.3 Foot-Controlled Model .................................. 127  
  6.1.4 Preliminary Evaluation ................................ 128  
    6.1.4.1 Apparatus and Implementation .................. 129  
    6.1.4.2 Participants ....................................... 130  
    6.1.4.3 Experimental Design ......................... 131  
    6.1.4.4 Experimental Task and Procedure ............ 132  
  6.1.5 Results ............................................... 133  
    6.1.5.1 Objective Measures: Task Completion Time ... 134  
    6.1.5.2 Subjective Measures: User Preference and Feedback 134  
  6.1.6 Discussion ............................................. 137  
6.2 Interfaces for Robot Arm Control ......................... 138  
  6.2.1 Haptic Control Interface ............................ 139
6.2.2 Gesture Tracking Interface .......................... 141

7 Conclusion .............................................. 143
   7.1 Summary of Contributions ............................. 143
   7.2 Design Challenges and Solutions ....................... 145
   7.3 Future Work ........................................ 148

A Experimental Material .................................. 151

References ................................................. 158
List of Figures

1.1 Overview of the rock breaker. ................................. 4
1.2 The ROM bin with a grizzly at the bottom. ..................... 5
1.3 Rock breaking by line-of-sight operation using a two-handed industrial control box. ................................. 6
1.4 The multi-tasking situation in remote rock breaking. .......... 8
1.5 Structure of the thesis. ......................................... 14

2.1 Evolution of natural interaction, adapted from [102]. ......... 20
2.2 Head tracking for FPS game, Source: [136]. ..................... 23
2.3 Eye gaze control for FPS game, Source: [76]. ..................... 24
2.4 Microsoft Surface ® multi-touch tabletop. Courtesy of Microsoft ................................. 25
2.5 iPhone and iPad. Courtesy of Apple ............................. 26
2.6 The Nintendo Wii, a gesture-based interface for gaming. Courtesy of Nintendo ................................. 28
2.7 The Microsoft Kinect, using the entire body as the gaming controller. Courtesy of Microsoft ................................. 28
2.8 General structure of a teleoperation system. ..................... 30
2.9 An example of using conventional user interfaces (e.g. joysticks, switches and keyboards) for mining teleoperation, Source: [59]. .... 33
2.10 Interactive PTZ camera control system using Wii remote and infrared sensor bar, Source: [54]. ................................. 35
3.1 A google display for head tracking. .............................. 38
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>The TrackIR system.</td>
<td>39</td>
</tr>
<tr>
<td>3.3</td>
<td>FaceAPI: real-time head tracking with a single webcam.</td>
<td>41</td>
</tr>
<tr>
<td>3.4</td>
<td>Using head tracking for dodging in FPS games, Source: [136].</td>
<td>42</td>
</tr>
<tr>
<td>3.5</td>
<td>Head rotation.</td>
<td>45</td>
</tr>
<tr>
<td>3.6</td>
<td>Head tilting.</td>
<td>45</td>
</tr>
<tr>
<td>3.7</td>
<td>Head flicking for pan.</td>
<td>45</td>
</tr>
<tr>
<td>3.8</td>
<td>Head flicking for tilt.</td>
<td>45</td>
</tr>
<tr>
<td>3.9</td>
<td>Front view of the experimental setup, Logitech webcam for head tracking (1), pelco camera (2), video stream from the remote camera (3), table soccer (4), covers for obscuring participant's direct vision (5), experiment participant (6), experiment assistant (7).</td>
<td>47</td>
</tr>
<tr>
<td>3.10</td>
<td>Back view of the experimental setup.</td>
<td>48</td>
</tr>
<tr>
<td>3.11</td>
<td>One hand switching between handles, the other performing camera control by keyboard.</td>
<td>50</td>
</tr>
<tr>
<td>3.12</td>
<td>Mean kicks for each camera control method.</td>
<td>51</td>
</tr>
<tr>
<td>3.13</td>
<td>Performance comparison based on computer game experience.</td>
<td>52</td>
</tr>
<tr>
<td>3.14</td>
<td>Performance comparison based on testing sequence (S1: Motion → Flicking → Keyboard, S2: Keyboard → Motion → Flicking, S3: Flicking → Keyboard → Motion).</td>
<td>52</td>
</tr>
<tr>
<td>3.15</td>
<td>Human head movements.</td>
<td>56</td>
</tr>
<tr>
<td>3.16</td>
<td>Our improved design of using head motion for remote camera viewpoint control.</td>
<td>57</td>
</tr>
<tr>
<td>4.1</td>
<td>Eyelink: a head mounted eye tracking system.</td>
<td>62</td>
</tr>
<tr>
<td>4.2</td>
<td>IBM BlueEyes Project prototype eye tracking system.</td>
<td>63</td>
</tr>
<tr>
<td>4.3</td>
<td>The FaceLAB eye tracker.</td>
<td>63</td>
</tr>
<tr>
<td>4.4</td>
<td>The Tobii T60 eye tracker.</td>
<td>63</td>
</tr>
<tr>
<td>4.5</td>
<td>Eye gaze control for large image inspection, Source: [2].</td>
<td>66</td>
</tr>
<tr>
<td>4.6</td>
<td>The gaze-driving robotic vehicle, Source: [127].</td>
<td>67</td>
</tr>
<tr>
<td>4.7</td>
<td>Gaze-controlled driving interface, Source: [127].</td>
<td>67</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

4.8 Design of gaze-driven remote camera control. 72
4.9 System diagram for the implementation of the gaze-driven camera control prototype. 73
4.10 Re-designed foosball table as a functional physical model. 79
4.11 Experimental setting: (1) participant, (2) remote camera, (3) screen view of video stream, (4) standard webcam for head tracking, (5) FaceLAB eye tracker, (6) gamepad as joystick based control interface, (7) re-designed soccer table, (8) covers for obscuring participant's direct vision. 82
4.12 Mean goals for each camera control method. 85
4.13 Mean kicks for each camera control method. 86
4.14 Mean results regarding questionnaire feedbacks (scale: 1 - strongly disagree, 2 - disagree, 3 - neither agree not disagree, 4 - agree, 5 - strongly agree). 87
4.15 Overall ranking by the participants for each camera control method. 89

5.1 Dual manual control mapping for both robot arm and remote camera. 101
5.2 Zooming by moving head towards the screen. 103
5.3 Relationship between the head lean position of a user (x axis) and the camera zoom level applied (y axis). 104
5.4 Remote setting: (1) Pelco ES30C camera, (2) Phantom Premium 1.5 as the rock breaker, (3) rocks, (4) hole for dropping rocks, (5) parking place for the arm, (6) ”nudge” function to simulate the ”fire” function. 106
5.5 User setting: (1) FaceLAB V4.5 eye tracker, (2) video stream from the remote setting, (3) Logitech Dual Action Gamepad. 107
5.6 Comparison of mean rocks sunk. 110
5.7 Comparison of mean nudges. 111
5.8 Mean numbers of rocks sunk for each operation period. 112
5.9 Mean numbers of nudges for each operation period. 113
LIST OF FIGURES

5.10 Mean results regarding questionnaire feedback (7-point Likert scale: 1 - Strongly Disagree to 7 - Strongly Agree) .................. 114
5.11 Overall preferences of the participants .................. 115

6.1 The ChairIO interface for First-Person-Shooter, Source: [8] ........ 122
6.2 Immersive traveling in a 3D virtual world by foot interactions on
a dance mat, Source: [1] .................................................. 124
6.3 Hardware construction of the WiiFitChair .................. 125
6.4 Coordinate system of the Wii Fit Balance Board (T: top, B: bot-
tom, L: left, R: right) .................................................... 126
6.5 Camera control configuration for using the FootMouse ........ 128
6.6 A participant uses WiiFitChair (left) and the foot-controlled model
(right) respectively to perform remote camera viewpoint control in
the experiment ............................................................. 130
6.7 (1): 4 groups of human face images with distinguishing features
with border colour (blue, red, green or black), gender (male or
female) and facial expression (smiling or non-smiling), (2): remote
camera, (3) the unique number on the forehead of each face ........ 131
6.8 Comparison of the mean task completion time .................. 133
6.9 Mean results regarding questionnaire feedback (7-point Likert scale:
1 - Strongly Disagree to 7 - Strongly Agree) .................. 135
6.10 Overall preferences of the participants .................. 136
6.11 Sensable Phantom Omni haptic device .................. 139
6.12 Task times for joystick and haptic control .................. 140
6.13 Design structure of Wiimote tracking .................. 141
List of Tables

3.1 User preference results from the subjective measures .......................... 53

4.1 Summary of user study results. ......................................................... 94
Chapter 1

Introduction *

Teleoperation has been widely applied in a variety of situations, ranging from space exploration (e.g., the National Aeronautics and Space Administrations (NASA) Mars Rovers [104]), inspection [115], robotic navigation [99], surveillance [64], underwater operations [20], rescue activities (e.g., searching for survivors at the Word Trade Center after September 11, 2001 [23]) to robotic surgery [7]. It has the promises and advantages of being able to provide replaceable surrogates for humans in hazardous or difficult working environments over long distances, potentially improving productivity and reducing costs. Regardless of whether the remote machine or robot is manually controlled by an operator, or semiautonomous, or even fully autonomous for some specific tasks, human observation, intervention, and supervision still play integral roles in these teleoperated systems [66].

Within the teleoperation model, the user interface is a fundamental and essential component that affects teleoperation performance significantly. Various types of user interfaces have been developed for different teleoperation settings, but directly controlling a robot while watching a video feed from the remote camera(s) remains the most common interaction form in teleoperation [47]. Therefore, in most teleoperation settings, an operator’s basic perceptual link to the remote

*Portions of this chapter have been published in [156] and [160].
environment is usually through a live video stream from a remote camera as the realistic foundation of situational awareness for the entire teleoperation activity.

In practice, multi-tasking situations are quite common as operators often have to control multiple devices simultaneously to complete operational tasks. A typical example can be controlling a mechanical robot and the motion of a remote camera at the same time. Using conventional control interfaces, such as joysticks, wheels, mice and keyboards, will result in frequently switching hands and attention between different control interfaces. This will distract the operator from concentrating on the control task, reduce the productivity of the entire process, and increase both the workload and the number of avoidable operational mistakes.

This thesis is motivated by a real-world industrial setting in mining teleoperation: Remote rock breaking, in which operators have to face such a multi-tasking situation. Instead of using conventional control interfaces, we explore several possible solutions by integrating natural human interactive information to design and develop new remote camera viewpoint control models. Natural interaction based interfaces allow people to interact with technologies, applications and systems as they are used to in interacting with other people in everyday life through movements, gestures, expressions, discovering the world by looking around and manipulating physical objects. In this research, we particularly investigate the use of natural head movements, eye gaze, weight-based body gestures and foot movements with developed prototype systems as alternative inputs for the camera viewpoint control in teleoperation.

A number of user studies have been conducted, in which we compare these camera viewpoint control models with conventional control interfaces. Several novel lab-based experimental models have been designed using the Functional Physical Model concept, which is used to simulate the real-world multi-tasking setting due to the difficulty of direct access to operators and industrial settings. Both objective and subjective measures were used in all user studies. From the results, we demonstrate our designed remote camera viewpoint control models are
either comparable to or an improvement over conventional control interfaces used in the experiments. Furthermore, we discuss existing issues, discovered problems and challenges of using natural interaction for remote camera viewpoint control. Finally, suggestions are made for further research in the area of importing more natural interaction based interfaces into teleoperation.

1.1 A Real-World Industrial Setting: Rock Breaking in Mining

In this section we introduce the real-world industrial rock breaking setting in iron ore mining. Mining iron ore in much of Australia is a complex process, and this description focuses only on the steps just prior to and subsequent to the operator actions of interest to us. We also address the multi-tasking problem in the development of a tele-robotic control system for this process [36] as an example application of our research scenario.

Iron ore is mined from the surface in this mine, with explosions used as the initial way to break up rock. The rock is then scooped up and loaded into giant 250-ton mining vehicles. These vehicles deliver their loads to a bin below which is a conveyor belt to a crushing machine to reduce the rock size.

As shown in Figure 1.1, the rock breaker on the mine site is a serial link manipulator arm with a large hydraulic hammer at the tip to break oversized rocks. The arm is installed at a ROM (Run of Mine) bin, where a number of horizontal bars (referred to as a grizzly) are fitted at the bottom in order to prevent oversized rocks from entering the crusher below (see Figure 1.2).

The actual rock breaking process is that those haul trucks dump their loads into the bin one by one at a certain time interval, in between an operator needs to break any oversized rocks stuck in the bin by controlling the rock breaker arm and firing the hammer at the tip. Using normal line-of-sight operation (see Figure 1.3) requires the operator to stand on the platform around the bin and
Figure 1.1: Overview of the rock breaker.
Figure 1.2: The ROM bin with a grizzly at the bottom.

manipulate the arm via two joysticks on a two-handed industrial control box. This process raises serious health and safety issues because of the harsh, dusty working condition in the mining environment and the risk of being close to the giant mining machinery.

The mining industry's long term plan is to teleoperate their mining operations [36], and achieve effective and safe remote control. Therefore, for the rock breaking process, a teleoperation system [36] has been developed to control the rock breaker arm at a long distance (over 1,000 kilometers), with a pan-tilt-zoom (PTZ) camera installed on one side of the bin as the basic visual feedback for the operator to accomplish the task.

The remote rock breaking process is shown in Figure 1.4. Instead of making an operator stand next to the bin, using a line-of-sight control to manipulate the rock breaker arm, the new remote setting allows the operator to have a desktop based teleoperation environment and live video as the visual feedback. The operator is required to break those oversized rocks stuck on the grizzly by operating a
Figure 1.3: Rock breaking by line-of-sight operation using a two-handed industrial control box.
two-handed joystick controller more suited to a seated posture rather than the original large industrial control box. As mentioned, operators have very limited time to break the rocks, as trucks arrive at short intervals (about 90 seconds). Since dumping a load raises a large cloud of dust, a water spray is used to settle the dust, which requires about 30 seconds to allow the operator have a clear view of the contexts of the bin. Therefore, the operator only has about 60 seconds to move the arm from its rest position, place it carefully onto a rock, break it by firing the jackhammer, and return the arm to the rest before next truck arrives.

When the operator is trying to break a rock, it is indispensable for them to have a close view (camera zoom-in view) of the target so that detailed information can be obtained to specify the spot on the rock for positioning the tip and firing the jackhammer. Otherwise, difficulties of positioning the tip on a proper spot of the rock would happen and slow the entire process, which could also result in avoidable operation mistakes. One serious issue of not having a zoom-in view in the tip positioning step is the arm would be bouncing on the rock and would easily be damaged after firing the jackhammer, if there was even a small gap between the tip and the surface of the rock. We verified the need for close in zoom by observation of and subsequent discussion with an experienced rock breaker operator.

It is practically impossible to mount the remote camera on the arm to couple the camera motion to the control of the remote robot like most telerobotic or vehicle settings for reducing the control complexity, as the camera would be quickly damaged when the jackhammer on the tip is fired to break a rock. Therefore, the remote camera is installed on the side of the bin with a zoomed-in view transferring the live video back to the operator. The operator has to use another joystick controller to control the camera motion to adjust the view of the target rock in order to complete the breaking spot inspection process, and then move on to the rock breaker arm control. This turns out to be a typical multi-tasking problem that requires operators to switch hands quite often between different
Figure 1.4: The multi-tasking situation in remote rock breaking.
control interfaces.

We take this typical industrial setting as a real-world example to address the multi-tasking problem for our research. Our natural interaction prototypes are developed based on the multi-tasking problem illustrated in this example. However, they could be applied in much wider situations with such multi-tasking problems rather than being limited to this particular mining teleoperation example.

1.2 Thesis Statement

Conventional interfaces have long been the dominant forms of user input for teleoperation, for instance joysticks, switches, wheels, mice and keyboards. They are relatively simple, sophisticated, allowing teleoperation to be a viable and profitable technique, which satisfies the basic client requirements for teleoperated systems with robustness and reliability. However, the ease of use, productivity, multi-tasking situations, hands-busy problems and frequently switching attention between tasks could be improved.

In this thesis, we build upon the concept of natural interaction [134] to investigate the possibility of introducing a user’s natural interactive information, such as head movements, eye gaze and body gestures as alternative user input mediums for camera viewpoint control in teleoperation. We posit that these forms of natural interaction can be used as essential elements to design new user interfaces to enhance the camera viewpoint control in teleoperation settings, which are able to improve task performance, provide intuitive control and more importantly offer effective solutions for the multi-tasking situation. Therefore, the thesis statement of this research work is:

"Natural interaction can be used in a practical user interface to enhance camera viewpoint control for teleoperation."

In this research, we explore the design space of interaction solutions that
use natural interactive information to develop advanced teleoperation interfaces, particularly focusing on improved remote camera viewpoint control models. Our goal is not to completely replace conventional user interfaces but to provide viable alternatives which could be used depending upon requirements of a particular teleoperation task, an operator's task ability and usability preference. Apart from the exploration on using head motion, eye gaze and gesture individually for the development, we also consider combinations of multiple interactive forms, such as using eye gaze in conjunction with head movements, which offer us a wider solution domain with more potentially effective possibilities.

We chose the realm of desktop interactions, since they are the most common settings used in teleoperation and also applicable to all types of computer users. In addition, recent computer vision technologies for desktop head tracking and eye tracking have been improved significantly, which are able to be potential alternative input modalities. The cost of these tracking systems may still remain as an issue, especially for eye tracking, however current technology trends indicate that low cost solutions would be possible in the near future.

1.3 Contributions

This dissertation presents a series of novel prototype models and corresponding user studies as a basis for integrating natural interaction into user interface design for camera viewpoint control in teleoperation. The major contributions presented in this thesis are:

- **Head tracking models**: Two remote camera control models were designed and implemented using two different sets of head gestures at the initial stage, named "head motion control" and "head flicking control". The results of the preliminary user study comparing these two models with another conventional keyboard control show that the head motion control is able to provide a comparable performance to the keyboard control but the
head flicking control is significantly worse. Using the user feedback, design improvements have been implemented on the head motion control which demonstrated better user performance and preference than conventional user interfaces in further experiments.

- **Gaze interaction model:** A novel gaze-driven remote camera viewpoint control is presented with an implemented prototype by integrating eye gaze information as an alternative control input, which follows a simple and natural design principle: "**Whatever you look at on the screen, it moves to the centre!**". A particular real-time gaze filtering (fixation detection) and smoothing approach is presented in detail to transfer the raw gaze data into corresponding remote camera control commands. The results of a user study are presented and discussed, comparing the gaze-driven model with a conventional joystick control as well as the improved head motion control.

- **Natural interaction model using gaze and head motion:** The design space of using gaze tracking in conjunction with head tracking has been further investigated. Apart from developing prototypes using either eye gaze or head movements, a new remote camera viewpoint control model is introduced by combining these two types of natural user inputs (named "natural interaction model"), using gaze information for camera pan and tilt control and head motion (leaning forward and backward) for zooming. A formal user study comparing this new model with an optimised manual control model and an autonomous tracking model is also presented.

- **Exploration of other models:** Another two remote camera control models which apply interactions of other parts of the human body are introduced. One is designed as a weight based gestural chair interface (named "WiiFitChair") with the advantages of being low-cost and interactive. It is built upon an existing popular game device (Nintendo Wii Fit Balance
Board). As to the other control model, the possibility of using foot interactions has been investigated. A Foot Mouse has been integrated as another prototype which uses simple foot movements to control the remote camera motion. The results of comparing these two models with a conventional hand-operated mouse control are presented.

- **Functional Physical Modeling for experiment design:** The *Functional Physical Modeling* concept for user experiment design is proposed with detailed modeling steps and applied examples. It is defined as a set of equipment which has been designed or selected to reproduce some specific properties of another set of equipment in a real world context for the purpose of experimentation, which can be particularly useful for cases where it is difficult to have direct access to the real-world settings.

- **Empirical user studies on developed prototypes:** In order to evaluate the usability of the developed prototypes, a number of empirical user studies on several lab-based settings have been conducted. Both user performance data and preference feedback have been collected through corresponding objective and subjective measures in various comparisons of different models. Data analysis has been conducted by using appropriate statistical methods. The results have been used as the essential feedback to assess as well as further improve these designed models. Also, these results presented can be used as practical references for designing or implementing similar natural interaction interfaces.

### 1.4 Organisation of the Dissertation

This thesis work was a collaborative project and completed at the Information Human Centered Computing (iHcc) Group in the Research School of Computer Science (RSCS) of the Australian National University (ANU) in conjunction with the Information Communication Technologies (ICT) Centre at the Commonwealth
Scientific and Industrial Research Organisation (CSIRO) with the support from the Minerals Down Under (MDU) National Research Flagship. Therefore, we use 'we' throughout the thesis for simplicity with contributions noted in section 1.3. The major scope of this research belongs to the field of Human Computer Interaction (HCI) but with the research perspective of importing natural interaction into interface design for teleoperation settings. This thesis does not deal with much computer vision or image processing research but takes the technological basis as granted. It also does not deal with much hardware construction but more on integration of the different parts of the prototypes.

This thesis consists of 7 chapters and 1 appendix (see Figure 1.5): an introduction, a literature review on natural interaction and related work done in teleoperation, 4 chapters on our major designed prototypes, user studies, results and relevant discussion, 1 appendix about used experimental material (e.g. consent forms and user questionnaires), and a summary of the entire work as the conclusion chapter at the end. The remainder of this dissertation is organised as follows:

In Chapter 2, we provide a literature review regarding the concept of natural interaction, the state of the art natural interaction related interface design and relevant sample applications (e.g. head/face interaction, gaze interaction, gestural interaction and voice interaction, etc), and previous work on interfaces for camera viewpoint control in teleoperation.

Chapter 3 - 6 present the developed remote camera viewpoint control models using different types of natural interaction with implemented prototype systems as part of this dissertation. Each chapter offers a self-contained section for a particular prototype design.

Chapter 3, on head tracking interaction, introduces two initial models: head motion control and head flicking control and a preliminary evaluation. From the experimental results and feedback, further improvements have been implemented for the head motion control.
Figure 1.5: Structure of the thesis.
1.4. ORGANISATION OF THE DISSERTATION

Chapter 4, on eye tracking interaction, presents a novel gaze-driven remote camera control model which follows a simple and natural design principle. A formal user study have been conducted, comparing the gaze control with a conventional joystick control, the improved head motion control. The Functional Physical Modeling concept is also introduced in the experiment design section of this chapter.

Chapter 5, presents a new natural interaction model combining gaze interaction with head motion with an empirical user study comparing an optimised dual manual control model and another autonomous camera tracking model in a simulated lab setting (another Functional Physical Model) with more key properties matched to the original real-world industrial setting.

Chapter 6, introduces another two models we have developed, one is a low-cost weight-based chair interface using the Nintendo Wii Fit Balance Board device as the hardware basis: WiiFitChair. The second one uses simple foot interaction by integrating a FootMouse. The evaluation in this section is the performance and user preference comparisons of these two models with a regular hand mouse on an object search task.

In Chapter 7, we conclude with a summary of the advantages and successes of these novel natural interaction models for the remote camera viewpoint control, reflect on lessons we have learned from the design process as well as the results from both objective and subjective measures, and make some suggestions for potential future directions.

Appendix A, includes relevant material used to conduct experiments, such as consent forms and questionnaires used as subjective measures to collect participants’ preference feedback.
Chapter 2

Background

This chapter presents the background material and related work relevant to this dissertation. It specifically looks at the definition of natural interaction, its history, current state of the art in this particular field, literature review of user interfaces for teleoperation and existing user interfaces for remote camera viewpoint control.

It should be noted that this chapter focuses on a general and comprehensive overview of the background, detailed analysis of related work that is relevant to each of our designed prototype is provided at the beginning of the corresponding chapter.

2.1 Natural Interaction and User Interface

For the past three decades, personal computing systems (i.e. the PC) have been incredibly widespread. Most of these systems were desktop-based settings. The user interface, the way or the "channel" a user "communicates" or "interacts" with a computer, has received a vast amount of attention. In fact since the creation of the Xerox STAR interface in the 80s, a huge amount of research effort has been spent on developing effective user interface techniques and metaphors. A variety of user interfaces have been developed for different types of applications,
however, using conventional user interfaces, such as a mouse and keyboard, to interact with standard monitor-based applications still remains dominant.

Nowadays, most people are familiar with using these conventional user interfaces for desktop computer systems. Due to the rapid development of diverse computing systems, these interfaces, techniques and devices are becoming inadequate, not very usable or for some cases not even able to satisfy the usability requirements. Therefore, apart from these conventional user interfaces, the design and implementation of more suitable and effective interfaces by integrating existing human interaction modes have become an essential research area and using natural interaction is the most important and active one.

2.1.1 What is Natural Interaction?

People naturally communicate through gestures, expressions, movements. The intention of conducting research in natural interaction is to invent and create systems or applications that can understand these actions or behaviors, allowing users to interact naturally with each other as well as the environment with intuitive engagement.

To formally define natural interaction is not easy at the current stage as it is a relatively new area for human computer interaction. However, there have been several relevant concepts introduced in the past research. In the late 1980s, Donald Norman first proposed the "natural mapping" concept in his book "The Design of Everyday Things" [106]. The notion of a "natural mapping" interface is that by taking advantage of physical analogies and cultural standards, it can lead to immediate understand for users. Another similar example was introduced by Timothy Poston and Luis Serra in 1994 [112], which was designed with the explicit goal of having an interface that would replicate how a user would perform dextrous tasks in everyday life.

In recent years, a few other corresponding expressions such as "natural user interface (NUI)" and "natural interactivity" have been used by many human
computer interaction scientists and psychologists in a vague or general way, or from a particular perspective. They are commonly used to describe something different from existing conventional interfaces.

The term "natural user interface" [113] is very much the first terminology regarding natural interaction. It is the common parlance used by designers and developers of computer interfaces to describe the emerging computer interaction methodology which focuses on human abilities such as touch, vision, voice, motion and higher cognitive functions such as expression, perception and recall. A natural user interface (NUI) seeks to harness the power of a much wider breadth of communication modalities which leverage skills people gain through traditional physical interaction.

"Natural interactivity" is another term with almost the same meaning as natural interaction but often used to demonstrate the difference from multi-modality [12]. Natural interactivity is multi-modal most of the time. However, a multi-modal system is not necessarily a natural interactive system. Multi-modality in a system merely signifies that users may, or must, exchange information with the system using several different input and/or output modalities [11].

The definitions of all these terms and expressions are very close or similar to natural interaction or can actually be recognised as the same. To better summarise these with a plausible definition, we take the notes stated in [134] as the basic foundation for natural interaction:

"Natural Interaction is defined in terms of experience: people naturally communicate through gestures, expressions, movements, and discover the world by looking around and manipulating physical stuff. The key assumption here is that people are meant to interact with technology as they are used to interact with the real world in everyday life, as evolution and education taught them to do."
2.1.2 History of Natural Interaction

The history of natural interaction is not long till recent days, but obviously it has become one of the most popular and active fields in human computer interaction and in the entire computing world. It can be tracked back to 2006, when Christian Moore [96] first established an open research community with the goal to expand discussion and development related to natural user interface technologies. Later, another conference presentation regarding natural interaction was given by August De Los Reyes [31] from Microsoft Surface Computing in 2008, which described the natural user interface (NUI) as the next evolutionary phase following the shift from the command-line interface (CLI) to the graphical user interface (GUI).

![Figure 2.1: Evolution of natural interaction, adapted from [102].](image)

As illustrated in Figure 2.1, in the early period of using command-line interface (CLI), people had to learn an artificial means of input, the keyboard, and a series of codified inputs, that had a limited range of responses, where the syntax of those commands was strict.

After the CLI, the widespread input device mouse enabled the graphical user interface (GUI), by which people could more easily learn the mouse movements
and actions, and were able to explore the interface much more. The GUI relied on metaphors for interacting with on-screen content or objects. The "desktop" and "drag" for example, being metaphors for a visual interface that ultimately was translated back into the strict codified language of the computer.

Much in the same way the graphical user interface (GUI) was a leap forward for computer users from command line interfaces, natural user interfaces in all of their various forms will become a common way we interact with computers. The ability for computers and human beings to interact in diverse and robust ways, tailored to the abilities and needs of an individual user, will release us from the current constraints of computing allowing for complex interaction with digital objects in our physical world.

2.1.3 State of the Art in Natural Interaction

Due to the rapid development of various sensing technologies (e.g. hardware sensors, vision-based sensing approaches), movements, expressions, gestures and sounds from different parts of human body can already be detected, processed and then used as alternative forms of inputs to design natural interaction based user interfaces for numerous types of control tasks and interactive applications. In the following part, we go through those major forms of natural interaction respectively to demonstrate the current state of the art in natural interaction.

2.1.3.1 Head and Face

The human head and face are rich sources for interface design of natural interaction. Both have served as modes of interaction and communication throughout history. People use head movements and facial expressions to express their emotion in the community, for instance, anger, fear, happiness, and also to "point" to an object to which they refer in their communications. According to these common capabilities, much research has been done in the investigation and development of such natural interfaces and interactive technologies.
CHAPTER 2. BACKGROUND

Head movements have been used for traditional pointing to offer an alternative control of the mouse cursor, which is very useful for people with hand and speech disability [13]. For example, in [53], a real-time vision system has been introduced, which performs cursor control by having users point their nose where they wish to place the cursor on a monitor screen. "hMouse" [50] is another head tracking driven camera mouse system, which provides alternative solutions for convenient device control. Other similar work can also be found in [98, 132, 133]

Recently, one of the most popular ways is to couple the virtual camera to a user’s head position in order to achieve a more realistic and immersive experience of perspective in virtual reality or visual gaming. For instance, in [136], head tracking has been integrated into a first-person-shooter (FPS) game "Bullet Time" to control the user’s viewpoint (see Figure 2.2). Another similar application with exaggerated head motions for game viewpoint control can also be found in [130]. Other relevant applications are head tracking based user interfaces for navigation in virtual environments, remote control of devices [6] and head gestures (e.g. "Nodding" and "Shaking") based perceptual interface [30, 97].

Furthermore, facial movements convey information about the emotions of the user [108], which can be used to detect confusion, user reactions, and intentions, and for purposes of surveillance and automatic video feed annotation such as in human behavior research. Computer vision based detection of facial expressions allows for an impressive perceptive interface and efficient video annotation tools, and can give emotional awareness to virtual agents [13, 18, 107].

2.1.3.2 Eye Gaze

The human eyes are both an input and an output channel. The usage of eye gaze can be found in many application areas. Predominantly, numerous experiments and studies have been conducted to investigate the user’s focus or attention on particular objects or tasks [29, 158], also for psychological-oriented intentions [63, 114]. For interaction purpose, most previous work of using eye gaze was
in assistive applications, aiding people with disabilities to control a computer in a hands-free manner [25, 72], similar to using head movements and facial expressions.

Apart from many approaches using eye gaze for traditional pointing, selecting and typing [5, 73, 81, 82, 89, 95, 118, 135, 142, 147, 148, 149], there have been a variety of other attempts to integrate eye gaze into the user interface design for different interactive models [26]. Tanriverdi and Jacob [128] presented an interaction technique that focused on combining features of eye movements and non-command based interactions particularly in virtual environments. It uses a histogram that represents the accumulation of eye fixations on each possible target object in the VR environment, which is able to provide a profile of the user’s "recent interest" in the various displayed objects. Similarly, there have been studies about integrating gaze-based interaction for video game control [123], using eye gaze as an alternative type of user intention to lead a group of virtual
agents/robots to accomplish cooperative tasks [51, 150], and experiments on the use of eye tracker for first person shooter (FPS) games (see Figure 2.3) [70, 76].

Especially in recent years, an obvious increase of research effort has been spent on developing eye gaze based interfaces or interactive models with the intention of not just being limited for disabled users. This implies that as part of human natural interaction, using eye gaze has become a very essential way for the integration of naturalness and intuitiveness.

2.1.3.3 Touch

The sense of touch always plays an integral role in human-being's sensing system. Early touch-based devices and applications were not wildly accepted because of the limitation of being single or dual touch. Since the breakthrough in the early 90s [137], multi-touch technology has been possible, offering much more types of touching interactions. Due to the advantage of naturalness by using multi-touch, various researchers and commercial companies started expanding
2.1. NATURAL INTERACTION AND USER INTERFACE

upon these inventions in the beginning of the twenty-first century and soon it became one of the most active fields with a large amount of applications [45] and technologies [60] being developed.

Microsoft introduced and started the development of their first table-top touch platform Microsoft Surface® [93] in 2001 (see Figure 2.4), which has the capability of interacting with both the user’s touch and their electronic devices. Similarly, Mitsubishi Electric Research Laboratories (MERL) began to design a multi-touch, multi-user system in the same year (DiamondTouch [33]) with the difference of being able to differentiate between multiple simultaneous users and it became a commercial product in 2008.

Figure 2.4: Microsoft Surface® multi-touch tabletop. Courtesy of Microsoft

Meanwhile, multi-touch truly became widely exposed and gained popularity through the association with Apple’s touching products. Apple was the pioneer to introduce multi-touch on a mobile device (iPhone [69]) and later they showed their invention of integrating multi-touch into portable devices by releasing the iPad [68] (see Figure 2.5).

Obviously small-scale multi-touch devices (e.g. iPhone and iPad) are rapidly increasing and becoming commonplace around the world. Clearly for using touch as a natural interaction mode, we can expect that in the near future, more intuitive, robust and customizable multi-touch solutions will become available, with interfaces and techniques that register multiple touch points and gestures.

25
2.1.3.4 Speech

The history of trying to develop speech (or voice) recognition systems can be tracked back to the early 50s, when people were attempting to make computing systems able to understand human-being’s linguistic communications. Even with such a long existence, speech interfaces (or referred to as Natural Language User Interface/LUI in human computer interaction) are still not often applied as they mostly suffer challenges to understanding wide varieties of ambiguous input and the inefficiency of the interaction [62]. The most common area of using speech interfaces is obviously within telephone communication where automatic or interactive voice response applications are commonly used.

Direct speech input for either control or interactive system design is often with a very limited vocabulary for simple input commands. For more complex applications or tasks, it is generally discarded. Notwithstanding, it seems to be more useful in multi-modal interfaces, combined with other input modalities or interactive styles. For example, Bolt [16] and Billinghurst [14] combined voice with gesture for graphical user interface interaction. It has also been combined with head movements or eye gaze [43]. In [37], another sample multi-modal interface using voice recognition and head tracking has been introduced particularly
2.1. NATURAL INTERACTION AND USER INTERFACE

for motor impaired users, which allows them to select distributed home appliances by their head direction and using simple voice commands for moving the selected target. Some newer interfaces like the one proposed in [83], add more intelligence to the interface that may also offer better acceptable interaction with more complex applications.

An important benefit of natural language interfaces is that users find them enjoyable to use, and are more satisfied with them than with many other interface styles [100]. A speech interface can be combined with other modes of interaction to broaden the range of interaction bandwidth. By taking advantage of as many interaction channels as possible, the efficiency and expressive ability of the interface can be increased. Therefore, as speech recognition becomes more robust, as natural language parsers are sufficiently powerful enough for restricted domains and as more processing power becomes available for the interface, speech-based natural language is becoming a more attractive method of interaction. By considering natural language as a viable interface style, the creativity of the design process need not be narrowed unnecessarily.

2.1.3.5 Gesture

Providing a complete overview of all gesture-related interactions is hardly possible, as it includes almost all commonly used interfaces and interactive modes. It ranges from the basic hand-operated devices (like mouse) to the entire body gestural interactions (e.g. wall climbing interface [35] and swimming interface [44]).

In fact, it has been the most active field of involving natural interaction based development in recent years. Particularly for gaming interaction, there have been a few "revolutions" on importing natural gesture interaction to improve gaming experience. The Nintendo Wii [138] (see Figure 2.6), a video game console released by Nintendo in 2006, soon gained its worldwide popularity as its distinguishing feature of using gesture-based interactions, providing unique engagement
in the gaming experience. The console uses a wireless controller, the Wii Remote, which can be used as a handheld pointing device and detects movement in three dimensions.

In 2010, Microsoft also released its new "revolutionary" gaming device: Kinect [77] (see Figure 2.7). It is a new sensor for the Microsoft Xbox 360 video game platform, which is able to provide a "controller-free gaming and entertainment experience" (or "You are the controller!" as announced by Microsoft). Based around a webcam-style add-on peripheral for the Xbox 360 console, it enables users to control and interact with the Xbox 360 without the need to touch a game controller, through a natural user interface using gestures and spoken commands.

Figure 2.7: The Microsoft Kinect, using the entire body as the gaming controller. Courtesy of Microsoft
2.2. USER INTERFACE FOR TELEOPERATION

Both Nintendo Wii and Microsoft Kinect (or even another early but similar gaming sensor EyeToy [40] from Sony) are gesture-based gaming devices, the success of these systems is mainly because the new interactive mode based on natural human gestures. Users can experience games in a completely new way different from traditional interactions, such as joysticks and gamepads. Many researchers and developers have been investigating possible applications that go beyond these systems' intended purpose of playing games. For example, Lee [85] takes the Nintendo Wii as an ideal platform for exploring a variety of interaction research concepts because of its impressive capability with a low cost and high degree of accessibility. The paper gives a detailed description of the technology inside the Wii remote, existing interaction techniques, what is involved in creating custom applications, and several projects ranging from multi-object tracking to spatial augmented reality.

Similarly, Kinect has also been combined with humanoid robot control, which has the robot respond to human gestures [126]. Another application called "depthJS" [32] integrates Kinect to allow users to control a web browser with hand gestures. Other ongoing projects like using Kinect to control PowerPoint and PDF presentations [3], and tracking groups of people in complete dark environments for video surveillance [34].

2.2 User Interface for Teleoperation

In this section, we briefly discuss the user interface issues in teleoperation. The remote camera viewpoint control will be particularly addressed due to its importance to the entire teleoperation activity. In addition, the reasons why we intend to apply natural interaction in the design of new user interfaces for teleoperation (especially for the remote camera viewpoint control) will also be summarised.
2.2.1 User Interface and Performance Issues

The benefits of using teleoperation are already apparent in numerous areas, ranging from space exploration, inspection, robotic navigation, surveillance, underwater operations and rescue activities. Due to the improving modern networks, users are able to be situated thousands of kilometers away to accomplish control tasks, allowing them to work under more convenient conditions or to use their skills at many different sites around the world. Particular examples like: web-based robot control [129], internet-based remote control for mining machinery [78], remote surgery and simulated collaborative surgical training using haptic devices [56].

Figure 2.8 illustrates the general structure of a teleoperation system. There has been a continuous long-term effort spent on researching and developing autonomous or semi-autonomous systems for a variety of teleoperational tasks, but human observation, intervention, and supervision are still integrally involved in these systems [66]. Within the entire teleoperation model, the user interface (usually contains displays and user control components) is a fundamental and essential component that affects teleoperation performance significantly [24].

Figure 2.8: General structure of a teleoperation system.

Due to various situations, remote robots can now be teleoperated or manipulated through a wide variety of user interfaces. For example, Fong et al. [48] developed a hand-held system for remote driving (called "PdaDriver"), which integrates a Personal Digital Assistant (PDA) system as the user interface for remote vehicles. It satisfies the requirements to minimise the infrastructure and
2.2. USER INTERFACE FOR TELEOPERATION

training, enable rapid command generation, and improve situational awareness. Another similar interface using a cellular phone for teleoperation has also been proposed in [120]. More complicated interfaces like the one introduced in [75], which uses multiple-panel displays with user control devices such as joysticks, wheels and pedals.

As discussed in [125], human performance issues in teleoperation activities often fall into two categories: remote perception and remote manipulation. The reasons why teleoperation activities tend to be challenging regarding performance are as follows [24]:

1. The performance is limited by the operator’s control skills and their ability to maintain situational awareness.

2. It can be difficult to build mental models of the remote environment.

3. Distance estimation and obstacle detection can also be difficult for the operator.

All these reasons are important guidelines for the user interface design of teleoperation in terms of performance. A detailed discussion regarding how remote perception and remote manipulation are affected by factors such as limited field of view (FOV), orientation, camera viewpoint, depth perception, degraded video image, time delay, and motion can been found in [24], also with reviews of corresponding techniques and technologies designed to enhance performance as well as to ameliorate potential degradations.

As suggested in [47]: to design a well-performing teleoperation system, the user interface must be as efficient and as capable as possible. All interfaces provide tools to perceive the remote environment, to make decisions, and to generate commands. Interfaces should attempt to maximise information transfer while minimising cognitive and sensorimotor workload. Moreover, an interface may also be designed to minimise training or to be user adaptive.
CHAPTER 2. BACKGROUND

On the other hand, we should also be mindful of that the involvement of the user operation will not be reduced as level of autonomy increases [47]. Even if a robot is fully autonomous, an operator still needs to know relevant information during task execution, which is particularly important when problems occur. Therefore, as the level of autonomy increases, user interfaces will be engaged more in monitoring and diagnosis rather than direct control.

2.2.2 User Interface for Mining Teleoperation

Particularly in the mining industry, safety is always an important issue. In order to prevent people from working under hazardous conditions or in difficult environments, teleoperation has been being increasingly adopted as an effective solution in mining industries (such as our research scenario mentioned in Chapter 1: the remote rock breaking example). Compared to the traditional line-of-sight control on mine sites, such a solution allows people to remotely control or manipulate complex mining machinery over long distances, with the merits of providing replaceable surrogates, reducing number of staff and being cost-effective [155].

Hainsworth [59] briefly discussed the requirements of user interface for teleoperation of mining vehicles and systems with the demonstrations of two teleoperated mining systems. It is clear that conventional user interfaces such as joysticks, switches, keyboards, and wheels, are still the major control elements used in mining teleoperation.

Figure 2.9 shows an example of using conventional user interfaces for mining teleoperation. Conventional user interfaces are relatively simple, sophisticated, allowing teleoperation to be a viable and profitable technique, which satisfies the basic client requirements for mining systems of robustness and reliability. However, operators often have to control multiple devices simultaneously to complete operational tasks, using a set of conventional control interfaces (like the sample setting in Figure 2.9), which will result in frequently switching hands and attention between different control interfaces. This distracts the operator from
2.2. USER INTERFACE FOR TELEOPERATION

Figure 2.9: An example of using conventional user interfaces (e.g. joysticks, switches and keyboards) for mining teleoperation, Source: [59].

concentrating on the control task, reduces the productivity of the entire process, increases both workload and the number of avoidable operational mistakes. Therefore, ease of use, productivity, hands-busy problem and frequently switching attention between tasks need to be improved especially for the user interface in mining teleoperation.

2.2.3 Remote Camera Viewpoint Control

Control of camera viewpoint plays a vital role in almost all teleoperation activities. Observations indicate that directly controlling a robot while watching a video feed from the remote camera(s) remains the most common interaction form in teleoperation [47]. The basic perceptual link between the user site and the remote environment is through a live video stream from a remote camera or a set
of cameras. Since teleoperation requires advanced immersive interaction, other types of feedback such as audio [121], tactile [22] and haptic [56, 57] displays have been available as useful supplements to traditional visual feedback but watching video streams is still the fundamental way for operators to obtain situational awareness from remote environments.

Using conventional interfaces such as joysticks to control remote cameras is the common way in most teleoperation tasks. Several alternative approaches have been developed based on some types of natural interaction, such as gestures. For example, Cohen et al. [27] proposed the possibility of using a set of circular oscillatory hand gestures to control a remote camera’s pan and tilt motion. In addition, due to the wide popularity of the Nintendo Wii in recent years as well as its advantage of low cost, another Pan-Tilt-Zoom (PTZ) camera control system using a Wii remote and a set of infrared sensors has also been described in [54] (see Figure 2.10). For these types of approaches, the basic intention is trying to provide more interactive or natural ways for the traditional remote camera control by alternative interaction modes. Nevertheless, they still require users to use attention and hands to operate, which are not applicable for the multi-tasking issue involved in our research.

2.2.4 Why Natural Interaction for Teleoperation?

As discussed in the previous sections, natural interaction has been recognised as the next generation for user interface in terms of its distinguishing feathers, such as naturalness and intuitiveness. It has been an obvious trend that natural interaction are being applied in numerous fields. Similar to other common user interaction fields, conventional interfaces are still predominantly used in teleoperation, but using natural interaction has the possibility to significantly improve the user interaction and productivity for teleoperation activities. We summarise the potential advantages of applying natural interaction in teleoperation (especially for remote camera viewpoint control) as follows:
Figure 2.10: Interactive PTZ camera control system using Wii remote and infrared sensor bar, *Source:* [54].

1. It may provide more natural and intuitive controls than traditional remote control.

2. It may be easier to use, with less complexity and present lower cognitive load for the control process.

3. It may require less training or even no training time and have lower cost.

4. It may be an effective solution for multi-tasking (e.g. hands-busy) situations as using different control modalities other than conventional hand operations/manipulations involved in teleoperation.

There have been several attempts to use natural interaction for remote camera control, such as using hand gestures and Wii remote introduced in the previous section. This implies the trend of applying natural interaction into teleoperation, and more relevant user interfaces and interactive techniques can be expected to become available soon due to these potential advantages.
2.3 Summary

This chapter is the background part to this dissertation. We specifically discussed the definition of natural interaction, its history, current state of the art in this particular field, including those major forms of natural interaction, such as head motion and facial expressions, eye gaze, touch, voice/speech and gesture.

The second part of this chapter gave a general literature review of user interfaces for teleoperation and existing user interfaces for mining teleoperation. The importance of camera viewpoint control to the entire teleoperation process has been emphasised with the introduction of several alternative remote camera control approaches using natural interaction. Furthermore, the potential advantages of applying natural interaction in the user interface design for teleoperation has also been summarised.
Chapter 3

Head Tracking Interface Design and Evaluation *

This chapter presents our design and evaluation of using natural head movements for camera viewpoint control in teleoperation. We first look at some related work about head tracking technologies and viewpoint control techniques using head movements. After these, our design of using head movements for remote camera viewpoint control will be presented, followed by a preliminary user study as the evaluation for our designed prototypes. After the detailed description of the experimental results, we discuss all the feedback and demonstrate the improvements made on the previous design.

3.1 Related Work

3.1.1 Head Tracking Technologies and Systems

Natural human head movements and gestures have served as a mode of interaction and communication throughout history. Responding to this common capability, much research has been done in trying to develop effective, robust and accurate head tracking technologies and systems to satisfy demand for building natural

*Portions of this chapter have been published in [151] and [153].
and interactive applications in the realm of human-computer interface design.

So far, various types of head tracking technologies have been developed. We can briefly classify these existing technologies according to the way head position is tracked into the following two main categories:

1. Sensor based head tracking.

2. Computer vision based head tracking.

3.1.1.1 Sensor based Head Tracking

The sensor based head tracking approach is fairly common. The typical configuration of this type of system comprises a set of sensors, which are required to be worn on a user’s head (e.g. head-mounted tracker), and another hardware device for detecting the position of the sensor, receiving the transmitted head data. It can be either connected to a screen based display, or goggles (see Figure 3.1) for visual feedback and interaction.

![Image of a goggle display for head tracking.](image)

Figure 3.1: A goggle display for head tracking.

There are a number of sensor based head trackers commercially available. TrackIR [101] (Figure 3.2) is a typical head tracking device currently quite popular amongst gamers, especially in the simulation community. This system consists
of a small infrared camera placed on top of the monitor and a prepared baseball cap with three IR reflecting strips. The camera tracks the position of these reflective markers on the user's head, and reports head position with 6 degrees of freedom. It offers a 120 Hz high sample rate and a \( 640 \times 480 \) raw sensor resolution. Head orientation can then be used as an input for many PC video games, for example, "fish tank VR", where a virtual world appears to be 3D as the view shifts depending on the angle of the user's current vision [123].

![TrackIR System](image)

**Figure 3.2**: The TrackIR system.

### 3.1.1.2 Computer Vision based Head Tracking

In recent years, much research effort has been expended on tracking and locating head pose, gestures and facial expressions from a video stream based on computer vision technologies. Compared with sensor based head tracking, this offers robust tracking quality, with more convenience and flexibility for the user as there is no need to wear any particular sensor devices, and less cost for the hardware as usually only a normal webcam is needed. The tracking frequency for this system
is about 50 - 60 Hz and it can track with as few as 40 pixels across the face (typically 2m from a VGA camera).

In the computer vision area, head tracking generally starts with 3D face detection by defining corresponding facial features. For example, using facial geometry is a major strategy to estimate the face location as well as head motion [15]. In addition, color information is another powerful cue for locating the face [65] and other methods such as the use of depth information [90], classification of the brightness pattern inside an image window [109]. Figure 3.3 illustrates a commercialized real-time face tracking technology: FaceAPI [41], which provides a suite of image-processing modules created specifically for tracking and understanding faces and facial features with 6 degrees of freedom for head tracking.

3.1.2 Head Tracking for Viewpoint Control

Head tracking is a key component in applications such as human computer interaction, person monitoring, driver monitoring, video-conferencing, and object-based compression. It has also been recognised as a natural choice for viewpoint control [19]. As we introduced in the previous section, the popular way of applying head tracking is mostly in virtual environments, coupling the virtual camera to a user's head position in order to achieve a more realistic and immersive experience for perspective or direction control.

Wang et al. [136] proposed head tracking as an augmented input for video game play. They introduced the way of integrating head tracking as another axis of viewpoint control in first-person-shooter (FPS) games, which allows a player to perform special interactions such as "dodging". A player’s view can be rotated (along the Z axis, see Figure 3.4) by a certain angle when they move their head sideways. Their user studies compared the game play with and without head tracking. From the results, they demonstrated that although there was no significant difference in players' performance, players showed their positive feedback of using head tracking which could effectively enhance the sense of presence,
Figure 3.3: FaceAPI: real-time head tracking with a single webcam.
CHAPTER 3. HEAD TRACKING INTERFACE DESIGN AND EVALUATION

role-playing and intuitive nature of the control.

Figure 3.4: Using head tracking for dodging in FPS games, Source: [136].

In addition, Teather and Stuerzlinger [130] introduced another system using exaggerated head-coupled motions to control the camera viewpoint, while performing 3D object movement tasks by a mouse. In their experiment, three exaggeration levels were compared. From the results, they suggested that there was some user preference for this type of exaggeration, but no significant differences by the experimental conditions were found, other than a learning effect. The advantage of a system like this is to allow the user to play the game with the primary input device, while simultaneously performing useful viewpoint control operations with their head.

Bernard [10] presented an interaction technique using head movements to control the location of a window viewpoint within its document space. In this work, head tracking is used in both a rate control interaction and a position control interaction. Two user studies built around two common GUI tasks: navigating in a two-dimensional document space, and moving an object from one place to another in a document were conducted. The results demonstrated significant improvements in task completion time after a short learning period of using head tracking for this window viewpoint control task.
More related to the control of real cameras, Nishikawa et al. [105] designed an interface, called "FAceMOUSE", which uses head movements to control the position of a laparoscope as an assisting device for solo surgery. In their system, a set of head gestures are used to control the camera functions of the laparoscope. For example, when the surgeon's head shifts parallel to the scope image plane, the laparoscope will perform the identical pan and tilt movements. This interface allows non-intrusive, non-verbal, hands-off and feet-off laparoscope operations. From the experiment of using this system in a real laparoscopic cholecystectomy on animals, they received positive feedback of the system being convenient for the surgeon as well as its potential applicability in clinical use.

3.2 Our Design of Remote Camera Control

In this section, we present our first design of using head movements for remote camera viewpoint control. The basic function of a remote camera is to pan or tilt. With various functional combinations, the operator can obtain flexible control of its movements. With the integration of head tracking techniques with these camera control functions, we propose two sets of simple head gestures as interactive methods for remote viewpoint control.

3.2.1 Head Motion Control

The first method (called "head motion control") operates according to natural human head motion. As shown in Figure 3.5, assuming initially that the user's head is directly facing the screen, when the user rotates the head to either left or right by a certain angle (shown as $\alpha$ in the figure), the camera will pan in the corresponding direction. It will keep panning the view along that direction until the user moves their head back to the original position. Figure 3.6 shows similar interaction for the head tilting. When the user tilts their head up or down by a certain angle (shown as $\beta$ in the figure), the camera will correspondingly carry

43
out the tilt function and not stop tilting until the head returns to the original position.

3.2.2 Head Flicking Control

The other set of head gestures is based on human quick head movements, called "head flicking control". The head flicking based interactive control for camera functions is mostly like a switch. When a user quickly rotates his head to either the left or right direction then moves back to the original position, we consider this to be a head "flicking" along the corresponding orientation, which appropriately turns on the camera to start panning along this direction. When the user flicks to the opposite direction, it will switch the camera movement off and stop at the current position. Figure 3.7 and 3.8 are the relevant geometrical displays of head "flicking" for both pan right and tilt up actions respectively.

3.2.3 Zooming Control

As for the zoom function, the general idea of this specific control is to operate according to the distance between the user's head and the screen. For instance, if the user wants to have a more detailed view of the current video streams, he may naturally lean closer to the screen, effectively suggesting that the camera conduct a "zoom-in" function, and vice versa. This is similar to the Head Butt Zoom approach introduced in [94] and the Lean to Zoom technique proposed in [61].

Please note that a more detailed description regarding the zooming control is given in Chapter 5, as it was not tested in the preliminary experiment presented in this chapter.
3.2. OUR DESIGN OF REMOTE CAMERA CONTROL

Figure 3.5: Head rotation.

Figure 3.6: Head tilting.

Figure 3.7: Head flicking for pan.

Figure 3.8: Head flicking for tilt.
3.3 Preliminary User Study

A preliminary user evaluation experiment was conducted to assess how well these two head gesture based methods could perform the control of a remote camera in a model of a real-world teleoperation setting.

3.3.1 Apparatus and Implementation

We integrated FaceAPI 3.0 [41] with a Logitech webcam [88] into our prototype system in VC++ that ran at 50 Hz on a PC for real-time head tracking. The system used a Pelco ES30C PTZ camera [109] to perform the head gesture based control for our study. A keyboard based method was also implemented to simply control the PTZ camera by using the four arrow keys on the keyboard.

The display was a 19" monitor with a resolution of 1280 × 1024 pixels for showing the video stream from the camera to the user. A half size soccer table was placed under the monitor with several covers attached on one side to obscure the user’s direct vision. Figure 3.9 and 3.10 show the experimental setup from front and back respectively.

3.3.2 Participants

A total of 10 university students and staff (8 male, 2 female) participated in this evaluation, ranging from 21 to 48 years old with a mean of 29.6 years.

All 10 participants were regular computer users with no previous experience in remote camera control. Four of them had some experience playing table soccer, and the rest had none. Most of the participants played computer games by using a keyboard occasionally (6 participants), one subject played quite often and the remaining 3 did not play games at all.
Figure 3.9: Front view of the experimental setup, Logitech webcam for head tracking (1), pelco camera (2), video stream from the remote camera (3), table soccer (4), covers for obscuring participant’s direct vision (5), experiment participant (6), experiment assistant (7).

3.3.3 Experimental Design

The experiment was conducted using a $3 \times 2 \times 3 \times 3$ within-subject design. Factors were control strategy (Head Motion, Head Flicking or Keyboard), table soccer experience (Never or Occasionally), computer game experience (Never, Occasionally or Often), and sequence of using three control methods. Since we had a small number of participants (10 participants), we decided to consider the sequence of using three control methods in our experiment as another testing factor. We chose three testing sequences (S1: Motion $\rightarrow$ Flicking $\rightarrow$ Keyboard, S2: Keyboard $\rightarrow$ Motion $\rightarrow$ Flicking or S3: Flicking $\rightarrow$ Keyboard $\rightarrow$ Motion) and allocated participants on each of them. Another reason why we did sequence selection was we also intended to see whether the conventional method (i.e. keyboard) would have
any performance effect on our head tracking methods.

Please note in Figure 3.9, the experiment assistant was seated at the back. His role was to gently and consistently return the ball to the participant when it was out of reach of their soccer handles. Thus, participants were essentially playing a one-player game.

Since the size of the entire play area was relatively small, we set the zooming level of the camera at a fixed value to only have a partial view of the field, leaving pan and tilt control to the participants. It effectively made the participants keep performing the control of the camera to find the ball throughout the whole experimental period, whenever the ball was out of the current area of vision.

### 3.3.4 Experimental Procedure

Participants were first given a short introduction (around 5 minutes) about the system, instructions on how to remotely control the remote camera by the two
3.4. EXPERIMENTAL RESULTS

types of head gestures and the keyboard, and what kind of task they would be required to accomplish in the experiment. After that, subjects started the experiment with the randomly selected sequence of using those three control methods. No pre-training period was offered. For each method, participants had 5 minutes to play the table soccer game, and the number of kicks they made was recorded for the performance measure.

Once the table soccer game under all the three conditions had been finished, the participant was asked to complete a short questionnaire in which they compared their experiences with different control methods across several criteria for the subjective measures, including easiness, naturalness and familiarisation time.

When conducting the keyboard based trial, there was no particular constraint for making the participant move both hands off from the two handles to the keyboard to adjust the view. As the control configuration of the keyboard was using only the four arrow keys, the participant could simply perform the camera control by using one hand pressing on the keyboard, leaving the other hand switching between two handles to kick (see Figure 3.11).

3.4 Experimental Results

3.4.1 Objective Results

A repeated-measure ANOVA analysis was conducted on the performance measure to study the effects of all the factors, i.e. control strategy, table soccer experience, computer game experience, and sequence of using three control methods.

The overall average kicks were 24.53. The control method factor had a significant impact on the final performance, $F(2, 22) = 5.6276$, $p < 0.05$. Participants performed best by using the keyboard ($M = 27.2$, $SD = 5.73$), the mean kicks by using head motion control was fairly close to keyboard control ($M = 25.9$, $SD = 8.29$), but the head flicking method had much worse performance ($M = 20.5$, $SD = 8.68$). Figure 3.12 shows the mean kicks for each control
Figure 3.11: One hand switching between handles, the other performing camera control by keyboard.
3.4. EXPERIMENTAL RESULTS

method.

![Bar chart showing mean kicks for motion, flicking, and keyboard control methods.]

Figure 3.12: Mean kicks for each camera control method.

Whether the participants had table soccer play experience did not have any significant effect on how many kicks they made in the experiment, $F(1, 22) = 0.0122, p > 0.05$. On the other hand, the factor of playing computer games using a keyboard turned out to have a significant impact, $F(2, 22) = 8.6814, p < 0.01$. Participants who often played computer games using keyboards outperformed subjects with only occasional experience or no experience through all three different control conditions (see Figure 3.13).

In addition, the sequence of testing these three control strategies for each participant was highly significant for the performance, $F(2, 22) = 15.8212, p < 0.0001$. Participants following the last testing sequence on our list (Flicking → Keyboard → Motion) performed the best using all the control methods, compared with participants using the other two testing orders. Subjects starting with keyboard control had much worse performance in general (see Figure 3.14).
Figure 3.13: Performance comparison based on computer game experience.

Figure 3.14: Performance comparison based on testing sequence (S1: Motion → Flicking → Keyboard, S2: Keyboard → Motion → Flicking, S3: Flicking → Keyboard → Motion).
3.4.2 Subjective Results

Table 3.1 illustrates the average scores participants rated for these three different remote camera control methods respectively according to their experience in the experiment. For the questions, we used a 4-point scale, rating from 1 (very difficult/very long) to 4 (very easy/very short).

<table>
<thead>
<tr>
<th>Average User Rated Point (out of 4)</th>
<th>Motion</th>
<th>Flicking</th>
<th>Keyboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1: How easy/natural did you feel in the experiment?</td>
<td>$M = 2.6$</td>
<td>$M = 1.9$</td>
<td>$M = 3.2$</td>
</tr>
<tr>
<td></td>
<td>$SD = 0.96$</td>
<td>$SD = 0.74$</td>
<td>$SD = 0.79$</td>
</tr>
<tr>
<td>Q2: How long did you feel to get used to the control method?</td>
<td>$M = 3.5$</td>
<td>$M = 2.9$</td>
<td>$M = 3.7$</td>
</tr>
<tr>
<td></td>
<td>$SD = 0.53$</td>
<td>$SD = 0.99$</td>
<td>$SD = 0.48$</td>
</tr>
</tbody>
</table>

Table 3.1: User preference results from the subjective measures

3.5 Discussion

Our objective results indicate that for this specific experimental setting, keyboards performed the best for most of the subjects. We believe this is due to the fact that all the participants were quite familiar with using a keyboard, and initially there was no training time for them to get used to the two head tracking control methods. The reason for requiring the subjects to immediately start performing the experiment was to test how well users could pick up the head tracking based remote control. It is clear that our "head motion" based design provides quite comparable performance to the most conventional device (keyboard) even without any training.

As we mentioned in the previous section, another reason might be because users were actually switching only one hand between two handles, leaving the other hand for keyboard control in the experiment, which did not cost them much extra effort to trace the ball and make kicks.

The "head flicking" strategy did not perform as well as the other head gesture
CHAPTER 3. HEAD TRACKING INTERFACE DESIGN AND EVALUATION

based control. From our observations, when users were conducting trials they had to flick their heads quite frequently in order to find the ball. This is because the size of the zoom in viewing area was relatively small so that it required users to adjust the camera view quite often, which made the entire control be annoying and inefficient.

We have found similar results in the statistical analysis according to users' computer game experience using keyboards. Users with more gaming experience performed better not only in keyboard control but also in both head tracking controls. This is probably because those subjects already had more game based interaction experience, and in this particular game-like environment they may engage in the task more easily.

The results of testing sequence analysis indicate that this factor had a highly significant impact on the subjects' performance. A few interesting points have been discovered by the comparisons. Participants following the second testing order (Keyboard → Motion → Flicking) had a decreasing trend on the performance \( M_{\text{Keyboard}} = 22.0 > M_{\text{Motion}} = 17.33 > M_{\text{Flicking}} = 13.67 \). In addition, the results of following the third sequence (Flicking → Keyboard → Motion) which used keyboard control between flicking and motion demonstrated another decreasing effect on these two head tracking controls \( M_{\text{Keyboard}} = 25.0, \ M_{\text{Flicking}} = 38.5 > M_{\text{Motion}} = 34.0 \). On the other hand, the results of conducting motion or flicking control first in the sequences produced performance which was significantly improved over using keyboard control first (i.e. \( M_{S1-\text{Motion}} = 26.0 > M_{S2-\text{Motion}} = 17.33, \) and \( M_{S1-\text{Flicking}} = 22.8 > M_{S2-\text{Flicking}} = 13.67; \) while \( M_{S3-\text{Motion}} = 38.5 > M_{S2-\text{Motion}} = 17.33, \) and \( M_{S3-\text{Flicking}} = 25.0 > M_{S2-\text{Flicking}} = 13.67 \).

We suggest that as all the participants were good at using keyboards, they might be highly locked into this very familiar interface through the whole experimental period. This over-trained skill would affect the learning process for subjects to get used to operating the new interfaces introduced subsequently in
the sequence. From our results, the use of keyboard control first actually depressed the performance of the two head gesture based methods.

The results of subjective measure are consistent with the performance measure. Keyboard control was ranked as the best, but users also suggested that the head motion control could be picked up very naturally without pre-training. Compared with these two methods, the head flicking control was the worst choice by users’ consistent dislike.

### 3.6 Learned Feedback and Improved Model

For our design, we follow an iterative process, therefore we can improve our model step by step. From the results of the preliminary experiment, we have seen the applicability of our simple designs for remote camera control. We list the feedback and points derived from the first experiment, which could be useful to improve the existing prototypes:

1. The head motion control offered a comparable performance to the keyboard control in the experiment; however, users can only use head motion to control pan and tilt function of the camera separately, they cannot use head motion to move the camera at a particular angle.

2. For the head motion control, users can only use either head rotation or tilting to perform the camera control. This design does not allow other natural head movements or gestures to control a camera, such as translations and head rolling.

3. The head flicking control turned out to be not a good design particularly for remote camera control. Participants felt it was annoying and difficult to control the camera motion by this method as they had to keep ”flicking” their head in the experiment. It also performed the worst as the results reflected.
CHAPTER 3. HEAD TRACKING INTERFACE DESIGN AND EVALUATION

According to these points of feedback from our preliminary study, we made improvements on the design, considering the fact that human heads can move with 6 degrees of freedom (DOF), including yawing, pitching, rolling and translation (see Figure 3.15). Please note that the pivot point (the "O" point in Figure 3.15) used here is the standard setting for designing interaction based on head movements. This is shown on the front of the head as usual, though the real pivot point on a human head is actually at the centre point inside the head.

![Diagram of human head movements]

(a) Yawing
(b) Pitching
(c) Rolling
(d) Translation

Figure 3.15: Human head movements.

Most of the time, it is difficult to recognise a user's head motion as a specific gesture, for instance, users may yaw and pitch at the same time. Therefore, instead of creating a set of particular head gestures as discrete commands for
camera control, the improved design only considers the distance between the neutral head position (the "O" point in Figure 3.16) and the moved position (the "M" point in Figure 3.16): when a user moves their head more than a certain distance (threshold) either horizontally, vertically, or diagonally, the camera will pan or tilt, or move along the same direction as the user's head move direction by conducting pan and tilt simultaneously (along the "θ" in Figure 3.16). We also applied the rate control mapping with a linear function gain to specify the moving angle and the velocity for the remote camera to carry out corresponding pan and tilt functions. This makes the camera movements proportional to the amount of the head movements performed by the user. We consider this to be appropriate for continuous control of pan and tilt functions as it allows users to have flexible head motion for the control of camera movements.

![Diagram of head motion](image)

Figure 3.16: Our improved design of using head motion for remote camera viewpoint control.
CHAPTER 3. HEAD TRACKING INTERFACE DESIGN AND EVALUATION

We also considered the physical limitation of the head movements: users cannot have a convenient view if they rotate or twist their head too much from the screen such as in the most common head tracking interaction used in virtual environments or FPS games, which simply couples the virtual viewpoint to a user’s head position. Our improved design triggers the camera to move when the user's head position has moved more than the threshold value from the neutral position, but instead of making the camera move exactly as much or proportionally as the head moves, the camera will keep moving the view along the direction until the user moves their head back to the neutral position (to a position smaller than the threshold). This reduces the degree of head movement required even for large amounts of camera movement.

The user evaluation of our improved head tracking design will be presented in next chapter, compared with our eye tracking design as well as conventional joystick control.

3.7 Summary

In this chapter, we presented an iterative process of designing head tracking based remote camera viewpoint control methods. We first introduced our approaches of using two different sets of human head gestures to control a remote camera with a preliminary experiment. The experiment we designed used a simple physical game analogue, modeling a multi-task environment for testing the users' performance through three different remote camera control strategies, including the head motion control, head flicking control and keyboard control.

From the results, we demonstrated the head motion control was able to provide a comparable performance to using a keyboard even without the requirement of pre-training, and the subjective measure of user's preference also indicated that the head motion control was a comparable and effective method for this remote camera control case. Furthermore, we found that the sequence of conducting the three methods is the most significant factor. The use of keyboard control first
depressed the success of using the other two head tracking methods. If the results of our experiment are maintained or consistent in longer term training and use setting, it would suggest a seemingly paradoxical training regime of using the least familiar and worst control method for initial training to enhance subsequent performance. This may warrant further investigation.

After the first round design and evaluation, the key points and learned feedback regarding issues of our first round designs have been listed as guidelines for the second step improvements. Considered all these facts on the previous prototype, improvements have been made with the demonstration of our improved design for head motion control.
Chapter 4

Eye Tracking Interface Design and Evaluation *

In this chapter, we focus on the design and evaluation of using natural eye gaze for camera viewpoint control in teleoperation.

We first look at some related work on eye tracking technologies, and gaze-based interactions and applications. Later, the design of using eye gaze for remote camera viewpoint control will be presented, followed by a formal user study as the evaluation for this prototype. After the detailed description of the experimental results, all the feedback and existing issues will be discussed.

4.1 Related Work

4.1.1 Eye Tracking Technologies and Systems

The eyes are a rich source of information for gathering context in our everyday lives. A user’s gaze is postulated to be the best proxy for attention or intention [146]. Using eye-gaze information as a form of input can enable a computer system to gain more contextual information about the user’s task, which in turn can be leveraged to design interfaces which are more intuitive and intelligent [148, 149].

*Portions of this chapter have been published in [52], [152], [153], [154], [156] and [159].
CHAPTER 4. EYE TRACKING INTERFACE DESIGN AND EVALUATION

The history of eye tracking can be tracked back to the late 19th century [144]. Early stage work includes using direct visual observation, mechanical techniques, reflected beam of light, suction caps and electro-oculography (EOG) to measure eye movements. Over the years, the approaches to eye tracking have evolved significantly and eye trackers now days have become less invasive, more flexible and accurate than their predecessors.

Similar to head tracking approaches, eye tracking technologies can also be classified into two categories:

1. Head mounted approaches.

2. Remote video based approaches (using computer vision techniques).

![Figure 4.1: Eyelink: a head mounted eye tracking system.](image)

Head mounted eye trackers are developed to fix the frame of reference for the eyes relative to the head movements. Some head mounted eye trackers can provide higher accuracy and frame rate than remote eye trackers as they can get a close up image of the eye by virtue of using the head mounted camera. For example, Eyelink [39] is a commercial head mounted eye tracking system which consists of three miniature cameras mounted on a padded headband (see Figure 4.1). Two eye cameras allow binocular eye tracking or selection of the subject's
dominant eye without any mechanical reconfiguration. An optical head-tracking camera integrated into the headband allows accurate tracking of the subject's point of gaze. This system is able to offer a 500 Hz high tracking rate with an accuracy of 0.5° of visual angle while most remote eye trackers can only provide a frame rate of about 50 – 60 Hz.

In recent years, remote video-based eye tracking approaches have become dominant with quite a few practical systems available. The basic method of
these approaches is measuring the motion of the centre of the pupil relative to
the position of one or more glints or reflection of infra-red light sources on the
cornea. These systems allow some range of free head movement, no need for users
to use a chin-rest or bite bar or "wear" a sensor in any way.

Figure 4.2 shows the BlueEyes project [67] at IBM Almaden developed remote
video based eye trackers which used infra-red illumination. Several commercial
systems [42, 84, 131] are also available, which use a similar approach and provide
non-encumbering, remote, video-based eye tracking (Figure 4.3 and 4.4).

For our research work, we used a FaceLAB (Version 4.5) eye tracker as shown
in Figure 4.3, which costs approximately $35,000. However, according to both
technology and economic trends, it is conceivable to have similar eye tracking
systems available with much lower cost in the near future.

4.1.2 Gaze Interaction and Applications

As part of human’s natural interaction ability, eye gaze has been recognised as
an augmented input medium or control modality in "advanced user interfaces"
[73]. The compelling reasons, advantages and motivations to design gaze-based
user interfaces or interaction models can be summarised as follows [74, 146, 147]:

1. It can be an effective solution for situations that prohibit the use of the
hands, for example, when the user’s hands are disabled (quadriplegic) or
continuously occupied with other tasks (such as the multi-tasking situation
in the remote rock breaking process).

2. Increasing the speed of user input, as clearly the eye can move more quickly
in comparison to other input mediums.

3. Reducing workload, repetitive stress, fatigue (nearly fatigue-free interaction
[116]) and potential injury caused by physically operating other devices.

4. Using eye gaze requires no training. It is natural for the users to look at
the object of interest, therefore the relationship between the control and the display is already well established in a user's brain.

5. Eye gaze is the best non-invasive indicator for a user's attention and intention. The eyes provide the context within which a user's actions take place.

6. Eye gaze works very well with hands in coordination, which can be a very compatible advantage for common desktop settings.

Therefore, numerous approaches, techniques, applications and systems using gaze-based interaction have been proposed and developed for various situations in the field of human-computer interaction (HCI). For instance, Jacob [73] investigated the usefulness of eye movements as a fast and auxiliary input mode with the introduction of several fundamental gaze-based interaction techniques, such as Object Selection, Continuous Attribute Display, Moving an Object, Eye-Controlled Scrolling Text, Menu Command and Listener Window. Zhai et al. [147] presented the MAGIC pointing technique. In this approach, the cursor is automatically warped to the vicinity region of the target where the user is staring and then they can use an additional pointing device like a mouse to manually finish the target confirmation or selection.

In addition, in order to resolve the classical "Midas Touch" [73] problem in gaze-based interaction, Kumar et al. [81] recently proposed a practical technique using a combination of eye gaze and keyboard triggers with a fluid look-press-look-release action, called EyePoint. Also, another recent approach of using modes to enable different types of mouse behavior to be emulated with gaze and by using gestures to switch between these modes, called Snap Clutch, has been introduced in [71].
4.1.3 Eye Tracking for Viewpoint Control

Especially when using eye tracking for viewpoint control, previous work was mostly in virtual environments and gaming. For example, in [123], eye tracking was used in a First-Person-Shoot (FPS) game where the player's viewpoint in the virtual world was controlled by the eyes. Similarly, Isokoski et al. [70] reported another experiment on the use of an eye tracker with a gamepad in first person shooter (FPS) games, where they compared three control conditions: (1) a traditional gamepad controller, (2) the combination of gamepad controlled moving and aiming with gaze, and (3) the gamepad controller used only for moving forward, with both the aiming of the weapon and steering of the movement done by gaze. There were no significant advantages for eye operated control according to the results, but they confirmed that eye tracker input can compete in killing efficiency with gamepad input in FPS games, which could be an effective approach to minimise the use of hand controls in FPS gaming.
Figure 4.6: The gaze-driving robotic vehicle, *Source:* [127].

Figure 4.7: Gaze-controlled driving interface, *Source:* [127].

Beyond controlling viewpoint in either virtual world or gaming interactions, eye tracking has also been used to control the viewpoint for other tasks. Adams et al. [2] introduced an approach of using eye gaze particularly for the inspection of large images. As illustrated in Figure 4.5, their method divides the screen into a central static region surrounded by a number of pan regions. When a user stares their eye gaze in one of the pan regions, the image will pan towards the corresponding orientation.

Tall et al. [127] recently constructed an experimental robotic vehicle which could be remotely driven by a gaze-controlled interface. In their system, the camera is mounted on the vehicle (see Figure 4.6) so that both can be controlled by simply looking at left or right on the screen to turn the vehicle into corresponding direction (see Figure 4.7). In their experiment, they investigated five different control inputs (*on-screen buttons, mouse pointing, low-cost webcam eye tracker and two commercial eye tracking systems*) for driving the robotic vehicle on a racing track. From the results, they found gaze control was similar to mouse control, which provides clear evidence that robots or vehicles can be controlled "hands-free" through gaze.

Another similar prototype that uses eye gaze for the control of a real-world robotic wheelchair and the viewpoint of the camera can also be found in [143].
Instead of dividing the screen into several regions or just left and right directions, this system allows the user to drive a wheelchair by simply looking at a set of command icons on the chair-mounted screen.

We chose eye tracking as we needed a control method which could be used for many hours a day in multi-tasking (hands-busy) and attention switch settings. The control method should therefore be natural to use for many hours to avoid physical harm to the user. People naturally move their eyes all the time and eye gaze is a natural signalling technique between humans [79]. Whether eye gaze is suitable for the remote rock breaking task is part of our investigation.

### 4.2 Design of Gaze-Driven Remote Camera Control

In this section, we describe the design of our gaze-driven remote camera control in detail. The basic input data for this approach is the real-time raw gaze coordinate value on the screen $P_i(x_i, y_i)$. After filtering the raw gaze points into fixations, we apply the "rate control" mapping with a linear function gain to specify the moving angle ($CAM_{angle\_current}$) and the velocity ($CAM_{velocity\_current}$) for the remote camera to carry out corresponding pan ($CAM_{velocity\_current\_pan}$) and tilt ($CAM_{velocity\_current\_tilt}$) functions. This approach follows a simple and natural design principle: "**Whatever the user looks at on the screen, it moves to the centre.**" which is similar to the "self-centering mechanism" in the "rate control".

Since human raw gaze points are inherently noisy [144], they are not suitable for direct application [73]. Two main forms of eye gaze are *fixations* and *saccades*. Fixations occur when a subject’s eye gaze pauses over informative regions of interest and saccades represent rapid gaze movements between points. For using gaze information as a form of real-time input to control a camera, it is more suitable to use fixations as smoothed data rather than noisy raw points to avoid
jerky camera movements.

We used a modified version of the Velocity-Threshold Identification (I-VT) fixation detection algorithm [118] to filter the raw gaze points from the eye tracker in real-time into fixations, as this method is straightforward to implement, runs very efficiently, and can easily run in real time. Instead of setting a velocity threshold, a gaze movement threshold was used in the modified version, in which two gaze points separated by a Euclidean Distance of more than a pre-defined value are labeled as a saccade. This is because the time for receiving each gaze point is the same, therefore it is not necessary to further calculate the velocity for each point as the distance value can be directly used. In the default implementation, we chose a distance threshold of 1 degree of visual angle.

We can break down the entire process into the following major steps (see Figure 4.8):

1. Processing the raw gaze data using I-VT algorithm to filter the noisy points (saccades), recognise the gaze fixation \( P_n(x_n, y_n) \) by calculating the centroid of the grouped non-noisy points [118]: \( P_1(x_1, y_1), P_2(x_2, y_2), ..., P_n(x_n, y_n) \).

\[
\begin{align*}
\bar{x}_n &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\bar{y}_n &= \frac{\sum_{i=1}^{n} y_i}{n}
\end{align*}
\]  

(4.1)

Fixations are usually in the range of 200 – 400 ms [118], so we used 200 ms as the time interval in our default implementation, which resulted in approximately 12 gaze points per round as the eye tracker we used is able to provide a 60 Hz tracking frequency. The value of \( n \) is the number of gaze points in the fixation category after filtering out the saccade points.
2. Calculating the distance $d$ and the angle $\theta$ between the current fixation position $P_n(x_n, y_n)$ and the centre of the screen $C_0(x_0, y_0)$.

$$d = |P_n C_0| = \sqrt{(x_n - x_0)^2 + (y_n - y_0)^2} \quad (4.2)$$

$$\theta = \text{atan2}(|y_n - y_0|, |x_n - x_0|) \quad (4.3)$$

3. If the current fixation $P_n$ is in the central area $C_0$ ($r_0$ represents the radius of $C_0$):

$$d < r_0 \quad (4.4)$$

the camera will remain at the current position.

If the current fixation $P_n$ is out of the central area $C_0$, the camera will start moving along the angle $CAM_{angle\_current}$ with its velocity $CAM_{velocity\_current}$:

$$\begin{align*}
CAM_{angle\_current} &= \theta \\
CAM_{velocity\_current} &= FG \cdot CAM_{velocity\_max} \\
FG &= \frac{d}{D}
\end{align*} \quad (4.5)$$

where $CAM_{velocity\_max}$ represents the maximum velocity of the camera. In our default implementation, $CAM_{velocity\_max} = 30^\circ/s$; $FG$ is a liner function gain calculated as a ratio between the distance of the current fixation to the centre $d$ and the maximal distance on the screen to the centre $D$. It is used to translate the ratio to the corresponding camera velocity proportionally, which means the further you look at from the centre, the faster the camera moves towards the centre.
4. The corresponding camera velocity on both pan $CAM_{velocity\_current\_pan}$ and tilt $CAM_{velocity\_current\_tilt}$ directions are calculated as follows:

$$
\begin{align*}
CAM_{velocity\_current\_pan} &= CAM_{velocity\_current} \cdot \cos \theta \\
CAM_{velocity\_current\_tilt} &= CAM_{velocity\_current} \cdot \sin \theta
\end{align*}
$$

(4.6)

The camera motion will keep following the user's current fixation direction, if its position is not in the centre area of the screen. The overview of the entire process is that wherever the user focuses their visual attention in the video stream, the camera will always bring that to the centre of the screen. Therefore, the user will not feel that they are actually performing much "deliberate control" of the camera movements.

4.3 Prototype Implementation

The prototype system contains two major parts: the user end and the remote camera site, in between can be a standard network connection. The overall system structure is illustrated in Figure 4.9.

At the user end, we integrated the FaceLAB (V4.5) [42] eye-tracking system (laptop version) into our prototype, which provides the real-time gaze tracking at a 60Hz frequency without the use of markers. This avoids the need to make the user wear any specialized devices, offering comfort and flexibility. Head mounted trackers can provide more accuracy and a higher tracking frequency but they are not comfortable to wear for long.

We used a Dell Precision Work Station with standard Window XP operating system installed as the main PC. The FaceLAB eye tracker was connected to the main PC through a local network for transferring the real-time raw gaze data. The FaceLAB Client Tools SDK was installed on the main PC, called by the gaze-driven camera control code for receiving the raw data from the local
Figure 4.8: Design of gaze-driven remote camera control.
Figure 4.9: System diagram for the implementation of the gaze-driven camera control prototype.
network. The control code translates the raw gaze data into corresponding camera control commands as we explained in the previous section, and sends the real-time commands to the remote camera through the external network. The laptop-based eye tracker shared the user screen for eye tracking on the main PC, as the user would only be watching the video stream from the remote camera on the user screen. The gaze-driven camera control code and other relevant software integrations were all implemented in Visual C++.

On the remote site, we used the Pelco ES30C [109] (the same mode of camera has been used in the real rock breaking setting) as the remote camera to be controlled in the prototype system with the capability to perform pan and tilt functions simultaneously. It was connected to the user end through an external network, transferring the live video stream back to the user and also receiving the control commands to carry out the relevant camera movements. The camera is able to provide 360° continuous pan rotation and a tilt range of +33° to −83° from horizontal.

Both gaze data processing and camera operation happen simultaneously as everything is being operated in real-time, i.e. the system does not operate the camera control on an iterative detect-gaze and move-camera process which would result in jerky movements for the camera. The camera motion is not able to match the original gaze tracking frequency. Therefore, fixations are used as smoothed inputs to synchronize the frequencies of sending and receiving control commands for the camera. Apart from the introduced 200 ms time interval for detecting each fixation, adding camera operation latency and network data transferring delay, the entire latency for the system is less than 250 ms, which has proved to be fairly tolerable in the user evaluation.

4.4 Evaluation

We wished to investigate how well the gaze-driven camera control could perform in a model of a real-world multi-tasking (hands-busy) setting, in comparison
to other control methods. Therefore, we implemented a joystick based camera control as an example of a conventional control interface in teleoperation by using a standard Logitech wireless gamepad.

For the joystick control, the user can control the remote camera motion by pushing the left joystick on the gamepad. The camera moves along the same angle as the user moves the joystick, and it will stop its movement if the user releases the joystick.

An empirical user study was conducted as the evaluation of the gaze-driven camera viewpoint control prototype. The experiment compared the gaze-driven approach with the joystick control. The improved head motion control introduced in Chapter 3 was also added in the experiment in order to further investigate the performance difference in terms of both objective and subjective measures for all three remote camera control methods.

4.4.1 Experiment Setting Modeling

As mentioned in the previous section, we had very limited access to the real rock breaker equipment, we used the concept of functional physical modeling [52] for our experimental task design, by which we could model many of the properties of the example setting by using another physical model.

Building a functional physical model is similar to "sketching" [21] or "prototyping" [124] a specific model for evaluation. Such concepts have been extensively applied in the design and building of various simulators with training purposes. Sample applications include virtual reality (VR) simulators for medical or surgical training [55, 56, 119], computer-based flight training devices [80], and immersive systems for military training [46, 91].

Our goal of this research is to improve teleoperation camera viewpoint control, which is motivated by the example real world setting of the rock breaker. Therefore, we designed the experimental setting to be similar in its key properties to improve the likelihood that our results can later be implemented in such real
world settings.

4.4.1.1 Functional Physical Modeling

A functional physical model [52] is defined as a set of equipment that has been designed to reproduce some specific properties of another setting in a real-world context for the purpose of evaluation. Generally, such a model would be appropriate for cases where any specific property is difficult to measure within the original setting, or there exists difficulties in access to either the original equipment or the real operators or even both, or some other unavoidable impediments to reproduce experiments with the original setting.

Here we use the terms 'original equipment' and 'original setting' to differentiate between the physical equipment and that equipment in its real world context. We use 'functional physical model' to be the equipment selected, found or constructed to model the 'original setting'. Occasionally, we will use 'new equipment' to mean the same equipment as the functional physical model, but without reference to our intent of modeling the original setting - we use the term so we can describe situations where perhaps someone else has done experiments with equipment which is similar to our new equipment. Normally we will use 'operators' to refer to the mining usage, being the people who operate the original equipment, and will normally use 'subjects' or 'experimental subjects' for people who will use our new equipment in experiments.

In experiments with a functional physical model, care must be taken to map the properties of the original equipment in the context of its original setting/operators to the new equipment/experimental subjects, and this mapping must be clearly enunciated to be able to make any conclusions from results on the model. Timescale is a property the mapping of which must be considered also.

Conclusions made from a functional physical model should be suggestive only, as any evidence arising from experiments has true validity only for the new equip-

76
ment, and is true for the original setting to the degree our functional physical model matches the original setting. For research purposes this limitation of being suggestive only is not a huge impediment. We prefer our results to have application to the specific task as well as to wider abstracted versions of that task, or application in general. A concomitant effect is however that if we discover something unexpected, we need to eliminate the possibility that the effect is due to the nature of our functional physical model, rather than the original setting. In some cases the driver for change and need for modeling is due to other changes, hence the setting to be used as the 'original setting' needs to be clearly identified.

4.4.1.2 Key Properties of Functional Physical Models

The Key Properties of Functional Physical Models are listed below:

- Functional physical models use some different equipment (selected, found or constructed) to model an original setting which is inaccessible, inconvenient or not amenable to reproducible experiment.

- The functional physical model can only be relied on to reproduce the specific properties which are clearly identified in advance.

- In constructing a functional physical model, it may be necessary model any or all of: a physical device, place of use of device, user characteristics of a regular or occasional user of the device, the opinion of a user of a device of said device (whether it is an accurate or fair opinion or not), and timescale of normal use of the device, etc.

- Neither a literal physical model or a literal virtual model has a priori more validity than a functional physical model to the original setting. Since we cannot easily access the original setting (or we would perform experiments directly), this introduces differences which we argue are best overcome by a functional physical model.
The applicability or validity of conclusions drawn from a functional physical model need to be argued in light of the planned mapping and any deviations identified from the experiment (whether deficiencies, or unplanned correspondences).

4.4.1.3 The Original Setting being Modeled

According to the modeling approach, we attempt to model the rock breaking process and perform experiments to determine a good or at least better interface for this setting. List of Properties of Original Setting are shown below:

1. The operators use two joysticks to control the robot arm. The left joystick is used more (controls x and y directions) while the right is used less (z direction).

2. Fine positioning of the robot arm is necessary, via operator controlling the pan-tilt-zoom camera to view the rock. We noted it was mostly on maximum zoom.

3. Gross positioning of the robot arm is done via the remote camera view.

4. There is limited time to break a rock, the opportunity to break rocks come fairly regularly (not all loads contain large rocks, but if two rocks arrive typically only one is broken between trucks). Except for the two rocks case, each rock breaking task is independent of the previous one, with little or no correlation between rock locations in the bin.

5. The operators are very engaged in their task and compete on a shift by shift basis with the other operators on fewer stoppages, related to efficient rock breaking.

6. The operator does not always have a good view of the rock to be broken. Strong sunlight/shadow boundary may occur in the bottom of the bin, which in this case may obscure the rock.
4.4. EVALUATION

We chose a game with which to construct our functional physical model primarily because of the competitiveness and engagement we observed among the operators in the original setting. We have observed that University students (who would be our experimental participants) become most engaged and competitive in games. Followed the choice we used in the previous head tracking experiments, we took the foosball game as the major task in this round user study as well, but with a significant amount of modifications on the original foosball table. The table part of our new equipment is shown in Figure 4.10.

Figure 4.10: Re-designed foosball table as a functional physical model.

Changes made to the foosball table are:

1. One pair of handles removed, and the right side goal is blocked by a sloping surface, to make this a one player game. A set of wheels with weights were attached, to make the handles centering similarly to joysticks normal spring-loaded self centering.

2. Surface contoured so that all points not reachable by the remaining men
are raised so the ball will not stop at such points.

3. Surface contoured with channels so that kicks which are 'too hard' tend to return to the back while kicks which are 'just right' progress towards the goal. The table also now gently slopes to the right, away from the goal, hence kicks which are 'too soft' will return to a rest position.

4. The stopping surface under the middle man on the left is quite slight. Thus it is rare for the ball to stop there. The left and right stopping places for the ball are positioned differently, the left ones being closer together.

The effect of the above changes is to create a single player game which retains the speed and excitement of foosball, but also allows the ball to be still.

List of Properties of Functional Physical Model are below, with descriptions of the match to the properties we selected as key properties of the original setting:

1. Instead of two joysticks, we have two handles which are rotated. The contouring of the surface means the ball ends up in the right part of the table much of the time, hence the right handle is used more. The minor property of self-centering in joysticks is reproduced by the wheels plus weights.

2. The surface contouring and channels on the table mean that precise kicks are only possible with a view of the ball and which man is nearby. The same model PTZ camera is used with a fixed close-in zoom.

3. There is no exact equivalent or need of gross positioning in our model. We allow no overview because this would break our model. If the experimental subject could see the whole table it would be obvious which handle and which man to kick with. There is still some overview or similar information available to subjects, as they can hear the ball, and often the ball collides with one of the men as it comes to rest. So, usually, subjects have an idea which handle at least they will need to use. In terms of the experiment, we
do not count kicks made when the ball is not in view on the screen, hence disallowing 'wild kicking' as a strategy.

4. There is no analog in our model of the limited time to break a rock. While there is limited time for each experimental condition, this is equivalent to a shift (period of work) in the original setting. The tendency of our subjects was mostly to kick too hard, which results in little correlation between the start and resting position of the ball. In our model, 1 kick models 1 rock broken.

5. The game is engaging for students, they reported enjoying the experiment.

6. The camera view is from the back, however we invert the video to correct for the left-right reversal. Subjects are able to use the camera view successfully to make kicks, just as operators manage to break rocks, hence this appears at least an acceptable analog of the original setting.

4.4.2 Experimental Setup

As to the actual experimental setting (see Figure 4.11), we used a standard 19" monitor as the major user screen with a resolution of $1280 \times 1024$ pixels, showing the video stream from the remote camera to the participant. The re-designed foosball table was placed under the monitor with several covers attached on the near side to obscure the participant's direct view of the foosball table so that the view of the setting was via the camera and screen only. Also, we integrated FaceAPI 3.0 [41] with a standard Logitech webcam [88] into our system as the head tracking control prototype.

4.4.3 Participants

A total of 30 undergraduate students (mostly first-year undergraduates, including a few mature age students) participated in this evaluation, including 23 male and
Figure 4.11: Experimental setting: (1) participant, (2) remote camera, (3) screen view of video stream, (4) standard webcam for head tracking, (5) FaceLAB eye tracker, (6) gamepad as joystick based control interface, (7) re-designed soccer table, (8) covers for obscuring participant's direct vision.

7 female, ranging from 18 to 32 years of age with a mean of 20.27 years old and \( SD = 3.33 \).

All participants were regular computer users (at least 2 hours per day) with video game experience of using a joystick based interface, but none of them had previous experience with either head tracking or eye tracking control interfaces, and none of them had experience with our re-designed foosball game.

Several participants wore glasses, the rest had normal vision without any correction, and their eye gaze could all be calibrated successfully.

4.4.4 Experimental Design

The experiment was conducted by using a repeated measures within-subject design so that all the subjects participated in all conditions of the experiment. The independent variable *camera control method* contained three levels:
1. Head tracking control.

2. Eye tracking control.

3. Joystick control.

The order effect was counterbalanced by using a Latin square.

4.4.5 Experimental Procedure

Participants took part in the evaluation individually. Prior to starting the experiment, participants were given a short oral presentation (around 5 minutes) about the user study. The context included an introduction to the system, instructions on how to control the remote camera by using the head tracking control, the eye tracking control and the joystick control respectively, and how to play the re-designed foosball game. The objective of their play was to score as many goals as they could during each camera control trial. All the participants were required to confirm an understanding of these introductions and the requirements of the experimental task.

After the completion of the oral introduction session, participants started the experiment directly, no pre-training period was provided before the formal experiment. For each control method, participants had 5 minutes to play the re-designed foosball game. An extra 3 to 5 minutes were spent on the calibration of each participant’s eye gaze before they started the gaze-driven control trial.

The video stream from the remote camera for each participant using different camera control methods were recorded respectively. In addition, their entire experimental period was also recorded by another video camera for further observations.

For the objective measures, the number of goals and kicks each participant achieved was recorded through checking against the video records. A kick is a purposive movement of a foosball man as controlled by the handles when it properly engages the ball by moving it some detectable amount. By purposive
we mean that if a man engages and moves the ball while it is not visible on screen then it is a random or accidental movement and not purposive. As we record the timing of kicks and record the screen view, purposive kicks are simple to determine.

Once the foosball game under all three camera control methods had been finished, we collected the participant's qualitative feedback on the prototype by using a questionnaire with a 5-point Likert scale, rating from 1 (strongly disagree) to 5 (strongly agree) and a short interview, in which they compared their experiences with different control methods across several criteria as subjective measures, including naturalness, required consciousness, distraction and time to get used to each control method.

4.5 Results

We report the results of the user evaluation through objective and subjective measures respectively. By conducting statistical analysis on the data from both of the measures, we demonstrate the comparisons of user performance and preference for all three camera control methods quantitatively and qualitatively.

4.5.1 Objective Measure Results: User Performance on Goals and Kicks

The major objective measures are based on the analysis of the number of goals and the number of kicks each participant achieved in the corresponding camera control trail.

A One-Way Repeated Measures ANOVA showed highly significant differences in scored goals between the three control methods, $F(2,58) = 9.29$, $p << 0.001$. Figure 4.12 shows the corresponding overall mean goals for each method. The post hoc tests revealed that participants using eye tracking control scored significantly more goals than using head tracking control ($p < 0.015$) and joystick
Figure 4.12: Mean goals for each camera control method.

Based control ($p < 0.01$). Although the overall mean goals using head tracking control ($M_{head} = 4.6$, $SD_{head} = 2.11$) is higher than the number using the joystick control ($M_{joystick} = 3.87$, $SD_{joystick} = 1.78$), there was no significant difference between these two methods in mean scored goals ($p > 0.05$).

The results of One-Way Repeated Measures ANOVA of mean kicks showed very similar results to the previous mean goals analysis (see Figure 4.13). Highly significant differences were found in number of mean kicks between the three camera control methods, $F(2, 58) = 15.85$, $p < 0.001$. Also, the post hoc tests revealed that eye tracking control significantly outperformed both head tracking control ($p < 0.01$) and joystick control ($p < 0.01$) in making kicks, but there was no significant difference in performance ($p > 0.05$) between using head tracking ($M_{head} = 24.23$, $SD_{head} = 4.17$) and joystick ($M_{joystick} = 23.7$, $SD_{joystick} = 4.24$). Thus, kicks correlate strongly with goals ($R(88) = 0.55$, $p < 0.01$), which relates to the rock breaker setting in firings of the jackhammer leading to successful breaking of the rock.
Figure 4.13: Mean kicks for each camera control method.

4.5.2 Subjective Measure Results: User Preference and Feedback from Questionnaire and Interview

The results of questions from the questionnaire regarding user feedback of naturalness, time to get used to the control, consciousness and distraction of using the three different camera control methods are depicted in Figure 4.14 which shows the mean results of all participants.

In addition, participants clearly indicated in the questionnaire that it was very annoying for them to use the conventional control interface, i.e. the joystick in the multi-tasking experiment as they had to frequently switch hands ($M_{annoying} = 4.1$, $SD_{annoying} = 0.92$).

All questions show significant results favoring the eye tracking control. The pair-wise comparisons between head tracking control and joystick control did not show any significant differences in questionnaire results through all the questions. From the mean comparisons in Figure 4.14, it can be seen that the first two ques-
Figure 4.14: Mean results regarding questionnaire feedbacks (scale: 1 - strongly disagree, 2 - disagree, 3 - neither agree nor disagree, 4 - agree, 5 - strongly agree).
tions regarding naturalness and time to get used to the control method show that participants ranked joystick control slightly better than the head tracking control. In contrast, the remaining two questions about consciousness and distraction show slightly better results of using head tracking control than the joystick, again these are not significant.

At the end of the questionnaire, participants were asked to state an overall ranking of the three camera control methods according to their experience gained in the multi-control experiment. The results are depicted in Figure 4.15.

The mean position for eye tracking control was 1.43, 2.17 for head tracking control, and 2.4 for joystick control. 20 out of 30 participants (66.7%) ranked eye tracking as the best method for the remote camera control in the multi-tasking experiment and 7 (23.3%) ranked it as their second preference. Head tracking was rated slightly better than the joystick with 6 (20.0%) participants ranking it as their first choice and 13 (43.3%) as the second choice. More than half of the participants (16/30 = 53.3%) ranked joystick control as the worst method in the experiment.

From the short interview conducted at the end of the study and the comments of the participants made whilst filling in the questionnaires, most of them felt that eye tracking control was an effective solution for the multi-tasking problem being modeled. In comparison to head tracking control, it was more convenient and flexible with less physical movement and less consciousness required to perform the camera control.

However, a few participants with much video game experience preferred to use head tracking control, as it provided a more interactive and immersive way for remote camera control and they felt more engaged in the table soccer game even though head tracking initially took a bit longer time than eye tracking for them to get used to.
Figure 4.15: Overall ranking by the participants for each camera control method.
4.6 Discussion

We further discuss the results shown above in a more detailed way. Instead of introducing more formal research findings of our user study, we would like to present observations across all the relevant information obtained at this point.

4.6.1 Further Discussions and Observations on Results

The objective results clearly indicate that for this multi-tasking experiment, our novel eye tracking based remote camera control method performed the best for most of the participants even without a pre-training period. Compared to head tracking and joystick control, more valid kicks and more goals were produced by using eye tracking control. It followed our general expectation that more kicks would lead to more goals, which applies to the other two control methods as well. It tells us that when using eye tracking control, participants could obtain significantly more opportunities to kick the ball and score more goals as the camera motion always followed their current visual attention, which provided an efficient and effective solution for acquiring situational awareness from the remote environment through the video stream.

The head tracking control prototype we designed did not perform as well as eye tracking control, but still provided very comparable results to joystick control, also without pre-training. This prototype could be considered as a possible approach for the multi-tasking problem as it offered a way to release a participant's hands from the camera control, but it still required participants to conduct plenty of physical movements of their head for adjusting the camera view, which is obviously not as efficient as just moving their eyes. In settings where there were multiple places an operator needed to look to make decisions this method may be more useful.

The subjective results and feedback from both questionnaire and short interview remain consistent with the objective results. Eye tracking control outperformed head tracking and joystick through all the criteria we selected. Especially
from the results of question 1 and 2 regarding the naturalness and time to get used to the control method, most of the participants clearly felt that they could pick up eye tracking control very quickly. This was our expectation that by importing such natural interactive information, it would result in obtaining easier remote camera control.

In addition, from the results regarding consciousness of control (Q3), we can see that there was not much consciousness or attention required for participants to control the camera motion by using eye gaze when simultaneously performing the other control task. Also, most of the participants directly mentioned that using this eye tracking control allowed them to pay significantly more attention on what they were doing with their hands so that they actually did not need to "think" much about how to control the remote camera. This effect clearly reflects back to the original notion of our eye tracking based design, which was to control a remote camera as little "deliberately" as possible.

As mentioned previously, the whole eye tracking prototype followed a simple design principle: "Wherever you look at the video, the camera will bring it to the centre of the screen." Therefore, there was no specific control configuration or mapping for participants to adapt to. In contrast, the camera view was automatically adjusted according to the participant’s current visual attention so that it apparently reduced the consciousness of one control task, and allowed more effort on the main control task rather than distracting participants from that task, which has also been reflected in the results about distraction in question 4.

The subjective comparisons between head tracking and joystick show that participants rated joystick control slightly better than head tracking control in naturalness (Q1) and time they spent to get used to the method (Q2). We suggest that regardless of the multi-tasking situation, participants might still feel that for performing camera control it was more comfortable for them to directly use their hands rather than using their head movements, as none of them had previous experience on any head tracking interfaces, but they all had experience of using
such a gamepad with joysticks in video games.

Also as several participants mentioned in their follow-up interviews, compared to the joystick, it initially took a bit longer for them to pick up the head tracking control, since they usually did not move their head as much as was needed for conducting direct control of a remote camera when looking at the video on a screen. They felt that using their head was performing more deliberate control of the camera than using eye tracking.

Furthermore, some participants commented that head tracking control might not be as practical and applicable as eye tracking in a real teleoperation setting, because users would more easily be fatigued and annoyed if they had to perform camera control by continuously moving their head for a long time, and our experimental period was fairly short (5 minutes), so that this effect was not really obvious.

On the other hand, participants rated head tracking better than joystick in consciousness (Q3) and distraction (Q4) as it helped participants resolve the multi-control problem, and using joystick control they had to spend more attention and effort on the extra control.

4.6.2 Existing Issues and Arguments

According to the experimental observations and the comments from the participants, the major issue with the gaze-driven camera control was the unreliability of the gaze tracking process. Several participants commented that it was still a bit sensitive and unreliable, as they were actually not able to have a completely free interaction. They noticed that when they occasionally moved their head direction along with the gaze unconsciously (it happened for most of the participants), the gaze tracking quality was reduced or sometimes the tracking would be lost if they moved their head a bit further away, which would directly affect the control quality of the remote camera. This was the major reason a sub-group of the experimental population ended up with the overall preference to the joystick.
control as reported by these subjects. Improvements in eye tracking hardware and software will reduce the effect of unreliability of the tracking process.

During the experiment as well as the post-experimental video record checking, we observed that a few participants attempted to score more goals by optimising their control behavior. Participants were using one hand on the joystick to control the camera, using the other hand swapping between the two foosball handles. This might be a difference in our functional physical model. That is, in the example real world setting, to carefully specify a breaking spot on a target rock required more precise positioning than required by our foosball model. In the rock breaking setting, when the proper camera view was achieved then the operator could carry out the final tip positioning. These are limitations of our work.

However, we believe and argue the mapping of our functional physical model to the rock breaking setting is quite plausible, as the re-designed foosball model matches many properties of the example setting. These include the similar control mechanism of the devices, similar operation process to complete the task, similar objective of the operation, competitive working condition, in a multi-tasking situation. Moreover, the foosball game is engaging for students and participants did report that they enjoyed the experiment. Thus there is some likelihood that our gaze-driven control could have similar benefits in the example real world setting as demonstrated in the evaluation.

4.6.3 Summary of the Results

Finally, we summarise both the quantitative and qualitative results from the user study in a compact way and visualize them in Table 1.

4.7 Summary

In this chapter, we presented the design and evaluation of using natural eye gaze for remote camera viewpoint control for the multi-tasking situation in teleopera-
<table>
<thead>
<tr>
<th></th>
<th>Head Tracking</th>
<th>Eye Tracking</th>
<th>Joystick</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>Good</td>
<td>Best</td>
<td>Worst</td>
</tr>
<tr>
<td>Participants’ Rank</td>
<td>2nd</td>
<td>1st</td>
<td>3rd</td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. immersive experience</td>
<td></td>
<td>1. natural control</td>
<td></td>
</tr>
<tr>
<td>2. hands free control</td>
<td></td>
<td>2. easy to pick up</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. not much consciousness</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. not much physical movement</td>
<td></td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. limitation on head motion</td>
<td></td>
<td>1. hardware may be expensive</td>
<td></td>
</tr>
<tr>
<td>2. not practical for long time continuous control</td>
<td></td>
<td>2. tracking quality could be improved</td>
<td></td>
</tr>
<tr>
<td>3. more physical movement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Solution for the Multi-Tasking Problem</strong></td>
<td>Possible</td>
<td>Good</td>
<td>Not</td>
</tr>
</tbody>
</table>

Table 4.1: Summary of user study results.
4.7. SUMMARY

We modeled the real rock breaking teleoperation by using Functional Physical Modeling, in our experiment a re-designed physical game analogue: foosball game was used as the major experimental task. Both head tracking and eye tracking camera control methods were compared with a joystick based control in a 30-participant user study, including both objective and subjective measures.

From the comparison of the results, we demonstrate that eye tracking control is able to provide a natural and efficient control for the movement of a remote camera, and participants performed significantly better than using either head tracking control or joystick control without pre-training in the experiment. There was no significant difference in performance between use of head tracking control and joystick control. The feedback and existing issues collected in the experiment have also been discussed with a final summary of all the results given at the end.
Chapter 5

Further Explorations of Camera Viewpoint Control *

After introducing prototypes using head tracking and eye tracking for remote camera viewpoint control in the previous chapters, we further explore design possibilities according to the lessons we learned from these prototypes.

In this chapter, we consider another 3 possible remote camera viewpoint control solutions: (1) dual manual control, (2) natural interaction (combining eye gaze and head motion) and (3) autonomous tracking, to resolve the common multi-tasking situation where an operator is required to control a robot and simultaneously perform the remote camera operation.

An empirical user study has been conducted on a model multi-tasking tele-operation setting as an another Functional Physical Model which simulates the entire rock breaking setting. The performance and user preference using these 3 camera viewpoint control models will be compared and discussed.

*Portions of this chapter have been published in [155] and [157].
CHAPTER 5. FURTHER EXPLORATIONS OF CAMERA VIEWPOINT CONTROL

5.1 Related Work

Using conventional control interfaces for the remote rock breaking process is a typical multi-tasking problem. As recommended in [24] that multi-modal interfaces provide a large range of interactions to improve traditional teleoperation control interfaces. Introducing additional modalities can supplement or augment an operator’s visual and control channels when heavily loaded. This suggestion has led to research developments not only in teleoperation, but also in many similar fields with standard user control or interaction settings, such as video gaming, virtual reality (VR) interactions. It is also the major motivation why we focused on developing such natural interaction based (head tracking and eye tracking) remote camera control interfaces for teleoperation settings.

Particularly for the multi-tasking situation, many researchers have investigated similar problems (e.g. manual input is not effective or both of the user’s hands are busy) with the intention of developing more intuitive control or interaction models that use alternative modalities other than hands. In [122], Sko and Gardner explore head tracking as an augmented input in first-person games with conventional keyboard and mouse control. They developed a set of interaction techniques and integrated these with a modern game engine. Moveover, Yamaguchi et al. [141] recently proposed another prototype using head movements to control a PTZ camera for online video chat, which allows users to do other tasks with hands simultaneously.

Similarly, as we discussed in the previous sections, human eye gaze has been recognised as an augmented input medium or control modality in advanced user interfaces. Numerous approaches and applications have been developed using gaze inputs in company with hand controls or in situations that the hand control is not possible. For example, Matsumoto et. al. [92] described a novel wheelchair system which has an intuitive interface using gaze and head motion of a user. The user needs only to look where they intend to go, and can start and stop by nodding and shaking their head. More recently, Yoo et al. [145] proposed another
novel attentive and immersive user interface based on gaze and hand gestures for interactive large-scale displays. The combination of gaze and hand gestures provide more interesting and immersive ways to manipulate 3D information.

All these related examples have demonstrated the effectiveness of combining different forms of input modalities for interaction or user control, especially for multi-tasking (hands-busy) settings. Therefore, based on the previous experience of developing either head tracking or eye tracking control methods, we extend the possibility of combining these two natural modalities together to develop and evaluate a more integrated remote camera viewpoint control solution in the following sections.

Additionally, beyond human-involved camera viewpoint control, there has been a long research track for developing autonomous camera operation approaches in computer vision and robotics. Cohen and Medioni [28] introduced a method to automatically detect and track moving objects using a PTZ camera for video surveillance. FLYSPEC [87]: another multi-user video camera system integrates manual and fully automatic control, which significantly assists users in remote inspection task. Obviously, fully automating cameras could reduce human control load for the operation process. We include an autonomous tracking model in our user study below.

5.2 Camera Control Models Considered

After a substantial amount of observation on the mine site, discussions with real operators and analysis of the user requirements, we propose three camera viewpoint control models as potential solutions for the multi-tasking situation in the remote rock breaking process.
5.2.1 Dual Manual Control - Conventional Interface

The first model is based on existing conventional manual control interfaces. The interface was selected to meet the user control requirements of this teleoperation process:

1. Both the robot arm and the remote camera should be operated by a single user control interface, i.e. dual manual control.

2. The control interface should be suitable for standard desktop use rather than the original two-handed industrial control box with a large size, which is not convenient for desktop based teleoperation settings.

3. The control configuration or mapping of the interface should be easy for the user to adapt to.

4. The control interface should be easy to obtain and maintain, also robust, portable and easy to complete integration and implementation for both the robot arm device and the remote camera.

We decided to use a standard Logitech Dual Action Gamepad [88], which satisfies all the requirements we had for the setting.

The manual control mapping for both devices is illustrated in Figure 5.1. The two thumb joysticks on the gamepad are used to control the robot arm. The left joystick moves the arm forward, backward and sideways, the right joystick controls the arm vertically, lifting up and down. The pan and tilt functions for the remote camera are operated by the four buttons on the right part of the gamepad, the Left and Right buttons for panning, the Up and Down buttons for tilting. Additionally, the two buttons on the front side of the gamepad are used to control zooming and the one below the "zoom out" button on the right side is used as a "fire" button to trigger the hammer.

Despite the complexity of this description these mappings are easy and similar to computer game control. Using this gamepad allows us to optimise the
Figure 5.1: Dual manual control mapping for both robot arm and remote camera.
manual control of two separate devices with a simple control mapping, which
minimises the transferring time of switching hand operations between different
interfaces. Furthermore, it uses a standard USB connection and it is a low cost,
very sophisticated interface which has been widely deployed.

5.2.2 Natural Interaction - Eye Gaze + Head Motion

The natural interaction model combines both eye gaze and head motion to control
a PTZ camera. From the previous performance comparisons between eye tracking
and head tracking control, we have found that eye tracking control significantly
outperformed head tracking control. Therefore, in this model, we inherit the
gaze-driven approach introduced in Chapter 4 to perform pan and tilt control for
a remote camera, which follows a simple and natural principle: "Whatever the
user looks at on the screen, it moves to the centre".

The camera zoom function is operated by the user’s head motion (see Fig-
ure 5.2), as small movements of the user’s head (approximately ±6 cm) can be
detected by the eye tracking device we had in real time.

The original version of the head motion for zooming control was similar to the
Head-to-Zoom in [2]: the zooming rate is proportional to the amount of movement
the user moves their head towards the screen or away from the screen and the
camera stops zooming in a small region around the user’s neutral head position.
However, this model did not perform well in our preliminary user test mainly
due to the zoom latency of the camera lens. We modified the model by adopting
another similar technique: Lean and Zoom [61], defining several discrete zoom
levels rather than using a continuous zoom scale. The camera zoom level is based
on the user’s current head position and the gain of the zoom can be adjusted
as desired. The relationship between the head position of a user and the camera
zoom level applied is shown in Figure 5.3. This improves the camera zoom latency
issue and also gives the user some freedom to move their head without activating
the zooming function, being more error-tolerant.
Figure 5.2: Zooming by moving head towards the screen.

5.2.3 Autonomous Tracking

The intention of an auto camera tracking system is that it automatically brings the view of the most important region of the teleoperation task to the operator. For our industrial setting, the operator mostly focuses their visual attention on the region around the tip when they are trying to specify the breaking spots on rocks and fire the hammer. Therefore, the view of the tip would be the region that the remote camera always tracks. In our model this is the tip of the robot arm.

An autonomous tracking approach has been mentioned in [36]. The design of the autonomous camera tracking system is to use hardware sensors on the arm to obtain the position data of the tip according to the pre-defined 3D coordinate system of the arm. Then the tip position data can be further geometrically transformed into the camera movement angle to make the remote camera carry out corresponding pan and tilt functions to track the tip in real time. This version
only considered camera pan and tilt automation.

For our model, we improved the design by adding an autonomous zooming function. When the operator is focusing on a target rock, a zoom-in view is essential. On the other hand, when the operator needs to make a "big" move of the arm, for example, trying to reach another rock far away from the current position, moving the arm to or from the parking place, then a zoom-out view would provide more visual information of the environment.

Therefore, a similar design to that of the head motion model has been implemented, in which the camera zoom rate is adapted to the distance the tip moved from the last position. Similarly, non-linear (e.g. sigmoid function) functions performed better to define the zoom rate than simple linear functions due to the high user tolerance for camera zoom latency as described in the previous section.
5.3 User Study

A user study was conducted to assess how well these three camera viewpoint control models could perform in a model multi-tasking teleoperation setting in terms of both objective and subjective measures.

5.3.1 Apparatus and Implementation

In this experiment, we still modeled the real-world rock breaker setting by building a physical control analogue and recruited university students as experimental subjects. Unlike the game analogue (foosball table) in previous evaluations, a Phantom Premium (V1.5) Haptic device [111] was used as a "small" equivalent of the rock breaker, which has a similar arm structure to the rock breaker device as well as very similar kinds of possible movements. We take this setting as another Functional Physical Model for the real rock breaker setting with a more obviously similar structure. It simulates the entire real-world process in almost exactly the same way. As we described in the previous section, the "small" robot arm was controlled by the two thumb joysticks on the gamepad for all three camera control models in the user study.

We modeled the entire rock breaker environment into a lab-based setting with similar properties (see Figure 5.4). The small robot arm is placed behind a wood board with a hole for dropping rocks in the centre and a parking place next to the board, which are similar to the setting of the real bin on the mine site. The same model of Pelco ES30C [109] camera as used in the industrial setting was installed on one side to be the remote camera for our user study.

In addition, since this haptic device based robot arm does not have the strength to break rocks, we implemented another function called "nudge" (see (6) in Figure 5.4) which makes the tip quickly move in a small circle (radius = 1 cm). This "nudge" function was designed to simulate the "fire" function for the real rock breaker tip. It is more similar to the industrial setting than just simply moving the arm to push rocks into the hole. It makes participants move the arm
Figure 5.4: Remote setting: (1) Pelco ES30C camera, (2) Phantom Premium 1.5 as the rock breaker, (3) rocks, (4) hole for dropping rocks, (5) parking place for the arm, (6) "nudge" function to simulate the "fire" function.

close to a target rock and "nudge" it into the hole, which is similar to the way that the real operator places the tip on a rock and fire the hammer, breaking the rock, the pieces of which fall through the grizzly.

For the natural interaction camera viewpoint control model, we used a FaceLAB V4.5 [42] eye tracking system to perform both real time eye tracking and head tracking at 60 Hz frequency without using any markers. The FaceLAB Client Tools SDK was used to implement the gaze filtering algorithm and the camera control code. The user control setting is shown in Figure 5.5.

All the software development was implemented in Visual C++ and installed on a Dell Precision Work Station with standard Window XP operation system on the user site. The display was a standard 19" monitor with a resolution of
Figure 5.5: User setting: (1) FaceLAB V4.5 eye tracker, (2) video stream from the remote setting, (3) Logitech Dual Action Gamepad. 1280 × 1024 pixels and shows the real-time video stream to the participant from the remote camera. A local LAN network was set up to connect the remote site and the user.

5.3.2 Participants

A total of 30 undergraduate volunteers (mostly first-year students) from a local university successfully participated in this user study, including 25 male and 5 female, ranging from 18 to 26 years of age ($M_{age} = 20.27$, $SD_{age} = 2.02$).

All participants were regular computer users (at least 2 hours a day) enrolled in a computing or IT related major with experience of using a gamepad interface. None of them had any previous experience on any eye tracking or head tracking interfaces, and none of them had any experience with our remote control setting.
Several participants had corrected vision but all the participants' eye gaze could be calibrated with the tracker successfully in the user study.

### 5.3.3 Experimental Design

The experiment followed a within-subject design therefore all the participants participated in all conditions of the experiment. The independent variable *camera control model* contained three levels:

1. Dual manual model
2. Natural interaction model
3. Autonomous tracking model

The order effect was counterbalanced by using a Latin square.

### 5.3.4 Experimental Task and Procedure

Participants took part in the study individually. The major task for the participants was to remotely control both the robot arm and the camera to "nudge" rocks into the hole, called "sinking" rocks by analogy with the game of billiards. Prior to starting the experiment, participants were first given a short verbal introduction about the system, a quick tutorial of how to control the remote camera by using the manual model, the natural interaction model and the autonomous tracking model respectively, and how to control the arm via the two thumb joysticks on the gamepad as well as the objective of their task, which was to try to sink as many rocks as they could during each testing trial. Participants were given a few minutes to experiment with all these control models, and then they were required to confirm an understanding of using these three different camera control models and the requirements of the experimental task.

After that, participants started the formal experiment. For each camera control model, participants had 3 minutes in total to sink rocks by nudging them.
Moreover, the entire 3-minute task was further separated to 3 operation periods (60s each), which was approximately the same length of time that an operator had in the industrial setting to handle one truck load dumped in the bin. After each operation period, participants were required to stop the nudging operation, and move the arm to the parking place. Meanwhile, an experiment assistant put rocks back on the board as a new dumped load, and then participants could bring the arm back and start the next operation round. An extra 3 to 5 minutes were required for calibrating each participant’s eye gaze before they started the natural interaction model trial.

During the operation periods, participants were required to use a zoom-in view of rocks while conducting nudging operations. This was to match the industrial operation in which a zoom-in view was vital. They could zoom out the view when they were looking for other rocks or moving the arm to or from the parking place.

The number of rocks each participant nudged into the hole and the number of nudges they made by using the corresponding camera control model were recorded as the major performance measures. Once the rock nudging task under all three camera control models had been finished, we collected participants’ feedback by using a questionnaire with a 7-point Likert scale, rating from 1 (strongly disagree) to 7 (strongly agree) and a short interview. The questions were derived from the IBM Computer Usability Satisfaction Questionnaire [86], in which participants compared their experiences with different camera control models across several criteria as subjective measures, including easiness, required consciousness, distraction and time to get used to the control and so on.

5.4 Results

5.4.1 Objective Measures: Rocks and Nudges

The major objective measures are the number of rocks sunk and nudges each participant achieved by using different camera control models.
CHAPTER 5. FURTHER EXPLORATIONS OF CAMERA VIEWPOINT CONTROL

A 1-way Repeated Measures ANOVA showed highly significant differences in number of rocks participants completed between the three camera control models, $F(2, 58) = 21.36, p < 0.001$. Figure 5.6 shows the relevant overall mean rocks sunk for each camera control model. The post hoc pair-wise tests revealed that using the natural interaction model participants sunk significantly more rocks than using the dual manual model ($p < 0.001$) and the auto tracking model ($p < 0.05$). In addition, using the auto tracking model participants performed significantly better than using the manual model ($p < 0.01$).

![Bar chart showing mean rocks sunk for manual, natural, and auto camera control models.](image)

Figure 5.6: Comparison of mean rocks sunk.

Similarly, the results of 1-Way Repeated Measures ANOVA showed differences between the camera control models were highly significant for number of nudges participants made in the experiment, $F(2, 58) = 23.75, p < 0.001$. The corresponding mean numbers of nudges for each camera control model are illustrated in Figure 5.7. After the post hoc pair-wise comparisons, we also found that participants made significantly more nudges by using the natural interaction model.
than the dual manual model ($p < 0.001$) and the auto tracking model ($p < 0.01$). Using the auto tracking model was also significantly better than using the manual model in the experiment ($p < 0.05$).

![Nudges](image)

**Figure 5.7:** Comparison of mean nudges.

We also separated the results of both rocks sunk and nudges to three phases for the three sequential operation periods which participants completed in the experiment. Figure 5.8 and 5.9 show the number of rocks sunk and nudges on average for each operation period respectively.

The overall trends of performance (including both numbers of rocks and nudges) for all three camera control models are increasing during the three operation periods. This reflects the general expectation that a participant’s performance would gradually increase at the early stage of a new task.

From the results of analysis, significant effects were found on camera control models for both mean rocks sunk ($F_{1st, 60s}(2, 50) = 21.64, p < 0.001$, $F_{2nd, 60s}(2, 54) =$
6.59, \( p < 0.01 \), \( F_{3rd.60s}(2, 57) = 7.8, p < 0.01 \) and nudges \( F_{1st.60s}(2, 59) = 23.1, p < 0.001 \), \( F_{2nd.60s}(2, 58) = 7.06, p < 0.01 \), \( F_{3rd.60s}(2, 57) = 10.2, p < 0.01 \) through the three periods. For all three period, participants using the natural interaction model significantly outperformed the manual model \( (p < 0.001) \). Using the auto tracking model was significantly better than the manual model at the 1st 60s period \( (p < 0.05) \) on both of the measures, but there were no significant differences on performance for the remaining two operation periods \( (p > 0.05) \).

5.4.2 Subjective Measures: Questionnaire and Interview

The results of questions from our questionnaire regarding user experience of using the three different camera control models are depicted in Figure 5.10 which shows the mean results of all participants.

The results of Friedman's tests showed significant differences on the user feedback in terms of camera control models for all the questions except for Question 4 and 8. Wilcoxon follow-up tests were used to analyse the results in more detail. For Question 1 \( (\chi^2_Q(1) = 6.92, p < 0.05) \), 2 \( (\chi^2_Q(2) = 6.94, p < 0.05) \)
and 6 ($\chi^2_{Q6}(2) = 26.91, p < 0.001$), participants rated the auto tracking model significantly better than the manual model ($p < 0.01$), but the results between the natural interaction model and the auto tracking model were not significant. Participants felt it took significantly shorter time to get used to the natural interaction model (Q2: $p < 0.05$) with much less attention (Q6: $p < 0.01$) than the manual model. As to Question 3 ($\chi^2_{Q3}(2) = 6.21, p < 0.05$) and 5 ($\chi^2_{Q5}(2) = 9.76, p < 0.01$), participants rated the natural interaction model significantly better than the other two models ($p < 0.05$), but the mean results between the manual model and the auto tracking model were not significant.

Highly significant differences were found in the remaining three questions, 7 ($\chi^2_{Q7}(2) = 24.54, p < 0.001$), 9 ($\chi^2_{Q9}(2) = 17.6, p < 0.001$) and 10 ($\chi^2_{Q10}(2) = 18.06, p < 0.001$). The mean results of both the natural interaction model and the auto tracking model were significantly better than the manual model for all the three questions ($p < 0.001$). In addition, the natural interaction model was rated significantly better than the auto tracking model for the last two questions (Q9: $p < 0.05$ and Q10: $p < 0.05$), but there was no significant difference between
### Figure 5.10: Mean results regarding questionnaire feedback (7-point Likert scale: 1 - Strongly Disagree to 7 - Strongly Agree).

<table>
<thead>
<tr>
<th>Statement</th>
<th>Manual</th>
<th>Natural</th>
<th>Auto</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. It was easy to use this camera control model.</td>
<td>5.07</td>
<td>5.73</td>
<td>6.03</td>
</tr>
<tr>
<td>2. It did not take long to get used to this camera control model.</td>
<td>4.8</td>
<td>5.67</td>
<td>5.9</td>
</tr>
<tr>
<td>3. It was simple to find the target in the remote environment using this camera control model.</td>
<td>4.53</td>
<td>4.9</td>
<td>5.77</td>
</tr>
<tr>
<td>4. I could gain enough visual information from the video stream using this camera control model.</td>
<td>4.8</td>
<td>5.77</td>
<td>5.33</td>
</tr>
<tr>
<td>5. I could obtain enough situational awareness for the teleoperation task using this camera control model.</td>
<td>4.57</td>
<td>5.2</td>
<td>5.9</td>
</tr>
<tr>
<td>6. I did not have to pay much attention on the camera control using this model when conducting the other task.</td>
<td>3.2</td>
<td>5.5</td>
<td>6</td>
</tr>
<tr>
<td>7. It was not distracting to use this camera control model when conducting the other task at the same time.</td>
<td>3.4</td>
<td>5.73</td>
<td>5.63</td>
</tr>
<tr>
<td>8. I felt the zooming function was easy to use.</td>
<td>5.1</td>
<td>5.3</td>
<td>4.83</td>
</tr>
<tr>
<td>9. I could quickly complete the task using this camera control model.</td>
<td>4.43</td>
<td>5.97</td>
<td>5.3</td>
</tr>
<tr>
<td>10. Overall, I am satisfied with this camera control model.</td>
<td>4.33</td>
<td>5.37</td>
<td>6.03</td>
</tr>
</tbody>
</table>

- **Manural**
- **Natural**
- **Auto**
them for Question 7.

At the end of the questionnaire, participants were asked to state an overall preference of the three camera control models according to their experience gained in the model multi-tasking experiment. The results are illustrated in Figure 5.11.

![Bar chart showing preferences](image)

Figure 5.11: Overall preferences of the participants.

The mean position for the natural interaction model was 1.53, 2.0 for the auto tracking model and 2.47 for the manual model. A majority of participants (19/30 = 63.3%) ranked the natural interaction model as their first preference and 6 (20.0%) ranked it as their second choice. The auto tracking model was rated better than the manual model with 7 (23.3%) ranked it as the best and more than half of the population ranked it as their second preference (16/30 = 53.5%). The manual model was rated as the worst by a majority of participants.
CHAPTER 5. FURTHER EXPLORATIONS OF CAMERA VIEWPOINT CONTROL

(18/30 = 60.0%).

From the short interview conducted at the end of the user study and the comments participants made whilst filling in the questionnaires, most of them felt that both the natural interaction model and the auto tracking model were effective solutions for the multi-tasking problem being modeled. In comparison to separate control interfaces requiring frequent hand switches, the dual manual model minimises the physical effort but participants still felt obvious distractions when they were carrying out the two control tasks at the same time in the experiment. They also commented that the natural interaction model offered a more intuitive way for remote viewpoint control than the other two models. Although the auto tracking model also worked well for the multi-tasking problem, participants felt that they lost some freedom of obtaining visual information from the remote environment as the camera could only be moved when they were moving the arm.

5.5 Discussion

The results of objective measures clearly demonstrate that for this model multi-tasking experiment, the natural interaction model significantly outperformed the other two camera control models. Compared to dual manual model and autonomous tracking model, participants nudged more rocks into the hole and certainly more valid nudges were produced by using natural interaction model. It followed our general expectation that more nudges would lead to more rocks to be completed. This tells us that using natural interaction model participants obtained significantly more opportunities to nudge rocks as the remote camera always followed their current visual concentration whilst operating the arm.

The auto tracking model provided a comparable performance to the natural interaction model, and participants performed significantly better than using the manual model. It offered a fully autonomous approach which effectively released any user involvement for the remote viewpoint operation in the experiment. How-
ever, this actually did not lead to the best performance in the user study as we expected, because it also reflects the fact that participants could not obtain as much situational awareness or visual information as by using the natural interaction model from the remote setting.

The subjective results and feedback from both questionnaire and short interview remain consistent with the objective results. Participants rated the natural interaction model significantly better than the other two models for most of the criteria we selected. As revealed in the results of questions regarding ease to use (Q1), time to get used to the camera control model (Q2) and required attention on the camera control (Q6), the auto tracking model was rated as the best as apparently it completely freed the participants from the control of the remote camera so that they were able to only concentrate on operating the arm to nudge rocks.

Nevertheless, from the results of questions about whether it was simple to find the target (Q3), whether users could gain enough visual information (Q4) and situational awareness (Q5) and efficiency of completing the rock nudging task (Q9), participants showed their preference for the natural interaction model rather than the auto tracking model. The major reason for this point was just as one of those participants with similar opinions in the interview commented that:

"Natural interaction model allowed you to look ahead and plan your next step move intuitively. However, auto tracking model actually did not allow you to look around or see further away as you did not have the control of the camera."

This also suggests to us that in complex teleoperation settings, beyond the current visual concentration for conducting the task, the way to improve an operator's capability of planning next step operation or predicting further situation might have a significant influence on the entire performance. Operating the remote camera is closely related to an operator's capability for planning and prediction, therefore most participants have significantly confirmed that using the natural interaction model offered them more visual information and situational
CHAPTER 5. FURTHER EXPLORATIONS OF CAMERA VIEWPOINT CONTROL

awareness than using fully autonomous model.

The results of the question regarding zoom function (Q8) are not significant, but a few interesting points were commented on by participants. Most participants directly commented that they really liked the interaction technique of using head motion for controlling the zoom function, since it makes the viewpoint control natural and intuitive. However, several participants with much experience of using the same model of gamepad for video games suggested that if only considering achieving the maximal performance, they would rather have the hand control of the zoom function on the gamepad, as they believed using head motion requires more physical effort than just pressing buttons on a gamepad. Similar opinions regarding the zoom function in the auto tracking model were also commented on by a number of participants. They explained that the auto zooming was actually a bit distracting and sensitive as it used only the distance they moved the arm, which sometimes resulted in frequently zooming in and out automatically. Therefore, they also suggested moving the zoom function to the hand control for the auto tracking model.

For our user studies, we simulated most properties of the mining process, which we consider to be a plausible experiment. This lab-based setting closely matches the real world setting, and it also provides full control of the experiment to test specific usability aspects on the control interfaces, i.e. the multi-tasking problem. In practice, to conduct such a formal user study on the real world mining setting will have a huge cost. Therefore, to design a model with the key properties of the real world setting will offer us an appropriate mapping to the real world setting, on which we are able to conduct user studies and obtain valid results with much less cost. In addition, the results would be useful at a more generic level rather than just for this particular mining scenario. In the experiment, participants were all novice users with limited training. From the results we can see that learning effect was still taking place, but it in fact offered us a fairer comparison of user performance under the same conditions than
5.6 Summary

In this chapter, we further considered three different camera viewpoint control models based on the experience obtained from previous prototypes and evaluations. The three new models include a dual manual control model, a natural interaction model using human eye gaze and head motion, and an autonomous tracking model.

A user study was conducted by modeling the multi-tasking situation based on a real-world industrial teleoperation setting (remote rock breaking). From the results of both objective and subjective measures on the model application, we reported the significant performance advantages and user preference for using the natural interaction model. The auto tracking model offered a comparable performance and the dual manual model was significantly worse. Our results were analyzed according to both quantitative and qualitative information collected.
Chapter 6

Other Models and Interfaces *

Apart from the remote camera viewpoint control models introduced in previous chapters, several other models and control interfaces have also been developed to further the research intention of this work.

We consider these additional models as a complement to this thesis since they are still at an early stage. The development intention behind these prototypes is not to just narrowly focus on performance comparison with other conventional interfaces but to further investigate the space of other forms of natural interaction into the user control interface for teleoperation settings.

6.1 More Models for Remote Camera Viewpoint Control

In this section, two more camera viewpoint control models will be presented with the results of a user evaluation: one is a low-cost weight based chair interface using the Nintendo Wii Fit Balance Board [139] device as the hardware basis (called WiiFitChair), the other model uses simple foot interactions by integrating a FootMouse [49].

Both models are additional attempts to explore other forms of natural inter-

*Portions of this chapter have been published in [57] and [103].
action (one uses weight-based body gestures and the other uses foot behaviors) for camera viewpoint control in order to potentially resolve the multi-tasking (hands-busy) issue we have been looking at for this thesis work.

**6.1.1 Related Work**

Previous research has already investigated various types of unconventional user interfaces such as chair-based interfaces and foot interactions, since such body-centered or torso-directed steering techniques enable natural proprioception and kinesthetic senses for interactions [17].

![Figure 6.1: The ChairIO interface for First-Person-Shooter, Source: [8].](image)

Beckhaus et al. [8] introduced ChairIO: a stool with two magnetic trackers installed on the sides of the seat, which allows users to use their body motion to perform translations and rotations in virtual environments. An informal user
study was conducted using this chair interface augmented by another game console gun interface compared with traditional keyboard and joystick controls in a First-Person-Shooter game (see Figure 6.1). Their results suggest that using ChairIO helps novice users to enjoy playing the game immediately, and it also offers a new form of gaming interaction for experienced players.

Since users navigate physically when interacting with information on large displays by frequently rotating their chair, Endert et al. [38] recently presented another chair-based interface: ChairMouse which captures this natural chair movement and translates it into large-scale cursor movement while still maintaining standard mouse usage for local cursor movement.

This chair-based interface is constructed by attaching an air mouse to a common office chair, which maps angular movement to linear cursor movement so that users can perform large-scale cursor movement by rotating the chair rather than continuously moving a regular hand-operated mouse. From the results of their experiment, they concluded that ChairMouse had a significant reduction of regular mouse movement in comparison to the traditional mouse only. In addition, users work more fluidly in the space, adopt better task strategies and make fewer mistakes by using this more embodied interaction interface.

Feet are also used in many real world tasks together with the rest of the body. They can easily be used as a supportive or augmented input channel in interaction with computers together with traditional hand-operated interfaces, e.g. mice and keyboards. Besides those conventional foot-controlled interfaces such as pedals, a number of other foot interaction models or devices have become available.

In [1], a body-centered interface is introduced by integrating a commercially available dance mat to perform foot controls for immersive environments (see Figure 6.2). From the results of a navigation study (i.e., travel and way-finding) comparing this foot-controlled interface with another hand-based interface, they demonstrated the effectiveness of using foot controls for immersive environments.

Similarly, another travel interface using foot controls on a dance mat has
Figure 6.2: Immersive traveling in a 3D virtual world by foot interactions on a dance mat, Source: [1].

also been proposed in [9], which is meant to be intuitive and allow hands-free interactions for the user.

6.1.2 WiiFitChair: Body Gestures + Head Tracking

To design a chair-based model, our initial intention was to develop a low-cost, intuitive, and easy-to-implement prototype. By considering these relative design aspects, we decided to construct the chair model by using a Nintendo Wii Fit Balance Board device as the hardware basis.

The Wii Fit Balance Board is a popular gaming device due to its body-centered interaction. It has also been demonstrated to be applicable as a new form of input device for a variety of virtual reality (VR) interaction scenarios [58]. As an input device, the Wii Fit Balance Board is a sturdy plastic panel that rests on four feet each of which contains a pressure sensor. This feature makes
the balance board an appropriate ergonomical equipment to have weight-based interactions. The hardware design of the chair-based model (named \textit{WiiFitChair}) was constructed by mounting the Wii Fit Balance Board under the seat of a regular office chair (see Figure 6.3).

![Figure 6.3: Hardware construction of the \textit{WiiFitChair}.](image)

The four pressure sensors give the balance board 4 degrees of freedom. The basic input data for this camera control approach is the real-time weight values ($W_{T,L}$, $W_{T,R}$, $W_{B,L}$, $W_{B,R}$) received from the sensors of the balance board (see Figure 6.4).

To process the balance board data, we apply the \textit{rate control} mapping with a linear function gain to specify the moving angle ($\theta$) and the velocity ($V_{\text{cam}}$) for the remote camera to carry out corresponding \textit{pan} ($V_{\text{cam-pan}}$) and \textit{tilt} ($V_{\text{cam-tilt}}$) functions. The mathematical model of the approach can be described as following steps:
Figure 6.4: Coordinate system of the Wii Fit Balance Board (T: top, B: bottom, L: left, R: right).

1. Calculate the weight differences on X and Y axis respectively:

\[
\begin{align*}
\Delta W_X &= (W_{T,R} + W_{B,R}) - (W_{T,L} + W_{B,L}) \\
\Delta W_Y &= (W_{T,L} + W_{T,R}) - (W_{B,L} + W_{B,R})
\end{align*}
\]  

(6.1)

2. If the weight difference value on either X or Y axis is less than the corresponding pre-defined threshold value ($W_{thre}$) then the camera will remain at the current position, otherwise move onto the next step:

\[
\begin{align*}
|\Delta W_X| &> W_{thre} \\
|\Delta W_Y| &> W_{thre}
\end{align*}
\]  

(6.2)

3. Specify the camera moving angle ($\theta$) and the velocity ($V_{cam}$):

$$\theta = \text{atan2}(|\Delta W_Y|, |\Delta W_X|)$$  

(6.3)
\[ V_{\text{cam}} = \alpha \cdot \frac{1}{1 + \epsilon^{-\lambda(t|\Delta W_x| + |\Delta W_y| - t)}} + V_{\text{cam\_min}} \quad (6.4) \]

Instead of using a simple linear function gain to specify the camera’s velocity, a sigmoid function is used to define the camera velocity proportionally to the weight difference values, which makes more smoothed and responsive camera movements.

In our default implementation, \( \alpha = 20, \lambda = 0.01, t \) equals to the threshold value \( W_{\text{thre}} = 30 \) which is adjustable in terms of a user’s preference, and \( V_{\text{cam\_min}} \) represents the minimal speed of the used camera.

4. Define the corresponding camera pan and tilt velocity using the following equations:

\[
\begin{align*}
V_{\text{cam\_pan}} &= V_{\text{cam}} \cdot \cos \theta \\
V_{\text{cam\_tilt}} &= V_{\text{cam}} \cdot \sin \theta
\end{align*}
\quad (6.5)
\]

As to the camera zoom function, the same head motion method introduced in Chapter 5 was adopted due to the design intention of an intuitive and hands-free camera control approach. The user’s head movements are detected by a webcam: leaning close to the screen leads to the camera zoom-in function and vice versa.

### 6.1.3 Foot-Controlled Model

For the foot-controlled model, we integrated a *FootMouse* into the remote camera control system. The *FootMouse* contains two components: a mouse pedal and a panel with foot-operated buttons and scroll, which gives users the ability to move the mouse cursor and click the buttons with their feet.

The remote camera control configuration using the *FootMouse* is shown in Figure 6.5. The camera pan and tilt function are simply controlled by moving the foot pedal. Users are able to move the remote camera at any angle when
they move the foot pedal by their right foot and simultaneously hold the mid-red button on the panel down by the left foot. The remote camera will stop moving when the pressed button is released. The zoom function is controlled by the scroll on the panel using the left foot: scrolling right zooms in and left for zooming out.

![Camera control configuration for using the FootMouse.](image)

Figure 6.5: Camera control configuration for using the FootMouse.

### 6.1.4 Preliminary Evaluation

A user study was conducted to assess how well these camera viewpoint control models could perform in exploring real-world objects in terms of both objective and subjective measures, which is similar to the ways operators have to explore the scene when conducting teleoperation activities. We designed a human face searching task for the user study, which has the advantage of being more compelling for our lab-based experimental setting.

Another conventional hand-operated mouse control method has also been implemented which adopted a similar control configuration to that used in the foot-
controlled model: controlling the camera pan and tilt by moving the mouse with the left button pressed; zooming in and out by sliding the middle scroll forward and backward.

Note that the intention of this preliminary evaluation for both WiiFitChair and the foot-controlled model is not to simply compete with the conventional mouse control model in terms of performance measures. In contrast, they are orthogonal and could be used in conjunction with any other hand-operated devices or control interfaces. Therefore, we believe that they are important complements and extensions to the development of new hands-free control interfaces for teleoperation.

6.1.4.1 Apparatus and Implementation

The software implementation of the WiiFitChair is based on the cross-platform WiiUse library [140] which handles the data communication between the balance board and a host computer via a Bluetooth connection. We further integrated the FaceAPI V3.0 [41] with a standard Logitech webcam to perform the real-time head tracking for the camera zoom control, which is able to provide a 60 Hz tracking frequency without using any markers. Both the FootMouse and the regular hand-operated mouse work with a standard USB connection, so no further software implementation was required. A Pelco ES30C camera [109] was used to conduct the remote camera control in the experiment.

All the software development was implemented in Visual C++ and installed on a Dell Precision Work Station with standard Window XP operation system. The display was a standard 19" monitor with a resolution of 1280 × 1024 pixels and shows the real-time video stream to the participant from the remote camera. A local LAN network was set up to connect the remote site and the user end. Figure 6.6 shows the experimental setting on the participant side.

On the remote site (see Figure 6.7), 16 human face images were semi-randomly separated into 4 groups. Each group has 2 male and 2 female faces with one smil-
Figure 6.6: A participant uses WiiFitChair (left) and the foot-controlled model (right) respectively to perform remote camera viewpoint control in the experiment.

ing and the other not smiling for each gender class. The 4 groups were arranged as a $2 \times 2$ structure and the face images within the same group distributed in a $2 \times 2$ pattern with different locations for the different face patterns.

All the face images were mounted on a standing screen, which was not visible to the participant. Each group also has a specific color on its border lines in order to distinguish it from other groups. On the forehead of each face, there is a randomly selected number which was to be reported by the subjects to uniquely identify this particular face when they were performing the face searching task.

6.1.4.2 Participants

12 undergraduate volunteers (mostly first-year students) from a local university successfully participated in this user evaluation, including 9 male and 3 female, ranging from 18 to 24 years of age ($M_{age} = 19.67$, $SD_{age} = 2.02$).

All participants were regular computer users (at least 2 hours a day) enrolled
Figure 6.7: (1): 4 groups of human face images with distinguishing features with border colour (blue, red, green or black), gender (male or female) and facial expression (smiling or non-smiling), (2): remote camera, (3) the unique number on the forehead of each face.

in a computing or IT related major with experience of playing video games. 4 of them had previous experience with games using the Nintendo Wii Fit Balance Board, the rest had no experience. None of them had any previous experience with foot-controlled mouse interfaces, and none of them had any experience on our remote camera viewpoint control setting.

6.1.4.3 Experimental Design

The experiment followed a within-subject design therefore all the participants participated in all conditions of the experiment. The independent variable Camera Control Model contained three conditions:

1. WiiFitChair model
2. Foot-controlled model

3. Regular mouse model

The order effect was counterbalanced by using a Latin square.

6.1.4.4 Experimental Task and Procedure

Prior to starting the experiment, participants were first given a short verbal introduction (around 5 minutes) about the system, instructions on how to remotely control the remote camera by these three control methods respectively, the structure of the face locations on the screen, and what kind of task they would be required to accomplish in the experiment. Participants were given a few minutes to experiment with all these control methods and they were required to confirm an understanding of using these control methods and the experimental task. For each of the control methods, subjects were asked to search for 3 different faces.

During the experiment, for each face searching case, a verbal instruction was given by an experiment assistant before the participant started the exploration, which was like:

"Could you move the camera view to the group with BLUE border lines, and try to find a SMILING FEMALE face?"

With those features in the instruction (including some global information via the border colors), participants could clearly distinguish the target face from the rest of the faces. After the participant successfully moved the camera to the location of the target face with its zoom-in view illustrated, they were required to speak out the unique number on the forehead of the face (see (3) in Figure 6.7), which was used to confirm the completion of this face searching task.

The instruction for next face searching task would not be given until the participant completed the previous one. There was no time break offered after the completion of experimental trials, participants were required to finish all the face searching tasks through one session, which took approximately 10 minutes per participant.
6.1. MORE MODELS FOR REMOTE CAMERA VIEWPOINT CONTROL

Once the face searching tasks under all the three conditions had been finished, the participant was asked to complete a short questionnaire in which they compared their experiences with the different control methods across several criteria with a 7-point Likert scale, rating from 1 (strongly disagree) to 7 (strongly agree) for subjective comparisons, and giving an overall rank for all three camera control methods.

6.1.5 Results

The results of the user evaluation are reported through objective and subjective measures respectively. We demonstrate the comparisons of the user performance and preference using different remote camera control methods in the face searching task quantitatively and qualitatively.

![Figure 6.8: Comparison of the mean task completion time.](image)
6.1.5.1 Objective Measures: Task Completion Time

The major objective measure is the analysis of the completion time each participant for the face searching task in the relevant camera control trial.

The results of a One-way Repeated Measures ANOVA show that the task completion time for the face searching experiment using the three remote camera control methods differed significantly, $F(2, 70) = 12.31$, $p < 0.01$. Figure 6.8 shows the overall mean task completion time for each camera control method. The results of the post-hoc pair-wise tests reveal that using the mouse control participants performed the best. They were able to complete the face searching task significantly faster than the other two camera control methods, the *WiiFitChair* model ($p < 0.01$) and the fool-controlled model ($p < 0.01$). In addition, the comparison between the chair model and the foot-controlled model showed very close performance, no significant differences were discovered in the analysis of the results ($p > 0.05$). This suggests the *WiiFitChair* was as good for the task as an unfamiliar mouse. The difference between the foot-controlled model and the hand-operated mouse should be due to familiarity.

6.1.5.2 Subjective Measures: User Preference and Feedback

The subjective measures are based on the user preference data and feedback collected from the questionnaire. The mean results for each question regarding participants' experience of using difference remote camera control methods are illustrated in Figure 6.9.

The results of Friedman's tests show significant differences on the user preference in terms of the camera control method for all the questions ($p < 0.01$). The results of the follow-up pair-wise (Wilcoxon) tests reveal that participants rated the mouse control significantly better than the other two camera control methods, the *WiiFitChair* model ($p < 0.05$) and the foot-controlled model ($p < 0.05$) for all the questions except Question 7. Similarly, the comparisons between the chair model and the foot-controlled model did not show any significant differ-
Figure 6.9: Mean results regarding questionnaire feedback (7-point Likert scale: 1 - Strongly Disagree to 7 - Strongly Agree).
Figure 6.10: Overall preferences of the participants.

ences in the results of these questions. Especially for Question 7, participants rated both the mouse and the foot control significantly better than the chair model ($p < 0.01$), but the results between the mouse and the foot-controlled model were not significant ($p > 0.05$).

The results of the overall participant preference regarding their experience for the three camera control methods are shown in Figure 6.10. The mean position for the mouse model is 1.17. The positions for the other two models are very close, 2.33 for the foot-controlled model and 2.50 for the chair model.

A majority of the participants ranked the mouse control as their first choice ($8/12 = 66.7\%$) and the foot-controlled model as the second ($8/12 = 66.7\%$). Most participants ranked the chair model as the last ($8/12 = 66.7\%$), but there are still a few participants showed their preference ($2/12 = 16.7\%$ as the best and
2/12 = 16.7% as the second) on the chair model over the other two models.

### 6.1.6 Discussion

From the results of the objective measures, participants performed the best by using the conventional hand-operated mouse control. In addition, the questionnaire results remain consistent to this point. The major reason for this could be the fact that the mouse is still the dominant input interface and users are very familiar with the hand-controlled model. However, our intention of including the mouse control model in the evaluation was not to compare it in terms of either performance or user preference as we aim to develop intuitive and hands-free camera control models. Obviously, using the mouse control was not a fair comparison to the other two models according to this research motivation. The purpose of the comparison was to take the most well-performing interface as a basic reference for users to provide feedback based on their user experience.

The *WiiFitChair* model and the foot-controlled model did not perform competitively to the mouse control in both the objective and subjective measures. However, all participants confirmed in their questionnaires that these two camera control models could be used as effective solutions for hands-busy tasks. A majority of the participants believed that their task performance could be improved significantly with more training for these two models, as very limited training time was offered before the experiment compared to the experience they had had of using a mouse.

For the chair model, participants were mostly able to handle the weight-based control effectively. The major issue on this model is the combination of the weight-based model and the design of using head motion for zoom control. As those participants commented after the experiment, they felt using head tracking was a very interactive design, which provided hands-free control for the camera zoom function. Nevertheless, leaning the head either forward or backward in order to activate the corresponding zoom in or out function would often result
in weight changes on the chair as their body would be moving with the head, which would lead to undesired and distracting camera pan or tilt movements to the user. This has turned out to be a significant issue which affected the task performance and the negative feedback revealed in the results of the subjective measures.

On the other hand, using foot-operated scroll to control the camera zoom function was more acceptable to the participants, which has been particularly reflected by the results of Question 7. Moreover, moving the foot pedal to control the camera pan and tilt function was acceptable but obviously not as efficient and accurate as hand operations by users. Several participants suggested the combination of using **WiiFitChair** and the foot-controlled zooming. Also, the fatigue issue was pointed out by a few participants as both two models require more physical effort than using a regular mouse according to their experience.

All these feedback, comments and discovered issues are the useful information we obtained from this preliminary evaluation, which needs to be taken into consideration not only for further design improvements on the existing models but also for future investigations on other similar interfaces.

### 6.2 Interfaces for Robot Arm Control

In this section, we introduce the interfaces particularly for robot arm control, which is related to the control of the giant mining equipment (rock breaker).

The first prototype we considered is using a haptic device as the user end to operate the remote robot arm, which is able to provide a real-time force feedback; the second interface we proposed is based on a gesture tracking approach which allows operators to use simple hand gestures to remotely control a robot arm.
6.2.1 Haptic Control Interface

A haptic interface can provide an operator with a sense of touching a "virtual" (or computer generated) object. The device used in this study was a Sensable Phantom Omni $^{[110]}$, which provides the sense of touch by pushing or pulling on the user's hand as they hold the end effector (see Figure 6.11). In a tele-operation scenario, the haptic interface can be programmed to reproduce some component of a real interaction occurring at some other place. The forces that the mining tool is experiencing can be detected, and these numerical values can be transmitted to the control centre. The user interface then can represent these forces via the haptic device, to the operator's hand. Because the transmission is computer mediated, those forces can be scaled or altered in any way, to suit the interface.

![Sensible Phantom Omni haptic device.](image)

With tele-operated mining equipment, the operator often needs to position large, multi-linked hydraulic arms. This positioning is a learned skill and it is often beneficial to locate it accurately on a particular part of a rock - be it a flat segment, a groove, crack or crevice. These features can be visible to a greater or lesser degree in an operator's view, depending on the quality of the viewing system installed, but any ability to feel the texture and irregularities of the rock
can enhance the operator’s ability to efficiently work the rock. We implemented the haptic interface to allow the operator to feel the rock that they are about to work on. The tip can be scratched along the rock’s surface, surface features can thus be felt as well as perhaps seen, with the aim of enabling a more accurate positioning of the tip.

To verify the use of this interface, we connected the software to a miniature robotic arm (refer to the model developed for the experiment described in Chapter 5, see Figure 5.4) in the laboratory to run a user study: comparing simple manipulation using the Phantom Omni against using a joystick-based gamepad. With the Phantom Omni the user moves their hand in 3 dimensions and the robotic arm copies the movement. With the joystick control, the user moves one joystick left-right and forward-back for similar arm control, and moves the other joystick arm for up-down control. Our aim was to test speed and ease of use in the task of pushing a rock into a hole. This is an analogy of the rock breaking task, the Phantom Omni is not capable of breaking even small rocks.

![Mean task time](image)

**Figure 6.12:** Task times for joystick and haptic control.

10 subjects, selected from computer studies students and workers, were run
through the task, half using the joystick first and half using the Omni first. The average time for the joystick interaction was 13.33 seconds with a standard deviation of 9.5 and for the 3D (Omni) interaction with haptics was 6.1 seconds with a standard deviation of 2.89 (see Figure 6.12). The 3D, haptic interface was also considered easier to use by 70% of the subjects. This agrees with [4] in that ideally, interactions with data should look and feel as if directly manipulating the data itself, without the need to pay much attention to the interaction itself. In a teleoperation context the "data" is the position of the tip of the robotic machine.

### 6.2.2 Gesture Tracking Interface

Inspired by the pioneering work introduced in [85] particularly about tracking fingers with the Nintendo Wiimote, we further investigated the possibility of using gesture tracking for the robot arm control. With the intention of building more interactive and low-cost control interfaces, we developed a preliminary prototype using a Wiimote with LED panels to perform hand gesture tracking. The design structure is demonstrated in Figure 6.13.

![Design structure of Wiimote tracking.](image)

For this tracking approach, the basic idea is by using an LED array and some reflective tape (e.g. put on fingers), we can use the infrared camera in the
Wiimote to track objects in 2D space. As shown in Figure 6.13, the IR camera of the Wiimote is capable of obtaining the positions of the two points in X and Y. Therefore, the position for moving the robot arm in the vertical plane can be obtained by calculating the middle of the two points. The horizontal direction then can be calculated with the distance difference between the two points. As the points and the view are fixed, when the two points come closer the distance between them will increase and vice versa.

This gesture tracking interface was integrated to the robot arm control system. An informal user study was then conducted with a few participants. Positive feedback was received regardless of performance measures, participants confirmed the interactiveness of this model which offers us potentials to further develop improved interfaces.
Chapter 7

Conclusion

The final chapter summarises the contributions of this thesis and synthesises the knowledge gained over the steps of this research by identifying the challenges in designing natural interaction based camera viewpoint control models and presenting solutions for addressing these challenges.

Potential future directions regarding importing more natural interaction strategies into the interface design and evaluation for teleoperation will also be discussed.

7.1 Summary of Contributions

Motivated by a real-world mining teleoperation example (Chapter 1), a typical multi-tasking setting (i.e. hands-busy situation), several novel camera viewpoint control models using natural interaction have been presented in this thesis. Detailed background reviews regarding natural interaction, its history and current state of art work, as well as previously related research about camera viewpoint control interfaces for teleoperation have been summarised in Chapter 2.

In particular, two new camera viewpoint control models integrating natural head movements were introduced in Chapter 3. An iterative design process was applied in this case. The results from the preliminary user study which compared
these two head models with a conventional keyboard control were taken as important feedback to designing improvements. The modified head control model showed a significant performance improvement in the further comparisons.

Chapter 4 looked specifically at the use of natural eye gaze as an alternative input modality for camera viewpoint control. A novel gaze-driven remote camera control model has been developed, following a simple and natural design principle: *whatever the user looks at on the screen, it moves to the centre*. In order to filter the noisy gaze data and provide smooth camera control, a particular real-time gaze fixation detection algorithm was applied. The *Functional Physical Modeling* concept for user experiment design was also introduced in this chapter with a sample model used in the user study in which the gaze control model was compared with the improved head control model and another conventional joystick control. This approach is useful for evaluation cases where it is difficult to have direct access to real-world settings. The results of the user study demonstrated the advantages of using the gaze-driven camera control model over the other two models.

Chapter 5 further presented a combination of gaze interaction with head tracking as a new natural interaction model for remote camera viewpoint control. An optimised dual manual control model and an autonomous camera tracking model were also developed and included in the user evaluation. A new Functional Physical Model was designed for the empirical evaluation, which more closely simulated the real-world sample teleoperation setting. The results showed that the natural interaction model significantly outperformed the other two camera viewpoint control models on both objective and subjective measures.

Apart from the head motion and gaze interaction models, the possibilities of exploring other forms of natural interaction for remote camera viewpoint control solutions have been discussed in Chapter 6. It proposed two more models: one was a low-cost chair interface using the Nintendo Wii Fit Balance Board to carry out weight based body interactions, the other one was a foot-controlled model by
using a foot mouse. These two models further investigated the design space of natural interaction. Moreover, some extra development regarding the robot arm control (i.e. the giant rock breaker in the sample real-world mining teleoperation setting) has been introduced in this chapter, such as using a haptic device with real-time force feedback and a hand gesture tracking based control interface.

From the results of a number of user studies, we demonstrated the remote viewpoint control models we developed were either comparable to or an improvement over conventional control interfaces used in the experiments. Particularly for the empirical user study comparing our Natural Interaction Model (combining eye gaze and head motion) with the Dual Manual Control and the Autonomous Tracking Model, the results indicate the advantages of using our Natural Interaction Model.

### 7.2 Design Challenges and Solutions

Designing camera viewpoint control models for teleoperation that incorporate natural interaction poses a number of challenges. According to the experience gained from the entire research process, we would also discuss the solutions addressing these unique challenges.

**How to find appropriate interaction forms?**

Natural interaction is based on the experience people have gained in their everyday lives. It includes a variety of interactive forms which can mostly be classified in terms of the use of different parts of the human body. It is an essential step to select the appropriate form(s) of natural interaction to be used in the design according to the existing problem domain (e.g. the multi-tasking/hands-busy situation in our real-world sample teleoperation setting) at the initial phase of the process. This thesis presented a thorough review of using natural interaction in user interface design, studied how people act or interact through these interaction forms, which could be
used for the camera viewpoint control design. The design of the camera viewpoint control models introduced in this research were all based on the existing multi-tasking issue addressed in the real-world teleoperation setting. Using a number of appropriate natural interaction forms such as head movements, eye gaze and body gestures to design camera viewpoint control models provides direct solutions to this particular problem.

Which sensors or sensing approaches should be used?
Various sensing or tracking technologies and systems are available for developing natural interaction based interfaces. It is difficult to find or construct a sensor that satisfies the user requirements of the design, is suitable for the environments or settings, has a balance between its sensing accuracy/data quality and the cost, and is easy to develop software implementations or integration also with ease of maintenance. The sensors and tracking systems used in this research have taken all these issues into consideration when designing the corresponding camera viewpoint control prototype. Both the head tracking camera control model (Chapter 3) and the gaze-driven control model (Chapter 4) used computer vision based tracking solutions rather than head mounted systems as these approaches were able to provide a more convenient user experience (i.e. marker-less tracking) with a sufficient tracking frequency (60 Hz for both systems). The other models such as the WiiFitChair model and the foot-controlled model (Chapter 6) were all constructed with the intention of being low-cost, easy to access and implement, robust and ergonomical.

How to interpret and use the sensor data?
Due to the limitations of the sensors/tracking approaches, the raw data generally contains noise or undesired information. Misinterpreting the initial sensor data will result in immediate failure of the entire system design. Therefore, it is necessary to transform the raw data into a format that can be used as appropriate user inputs with useful information, which
may require the development of specific algorithms or mathematical models. For example, the gaze-driven camera viewpoint control model proposed in Chapter 4 used a particular gaze filtering and smoothing approach. By applying a fixation detection algorithm, fixations are interpreted from the raw gaze data in real time and further transferred into remote camera control commands for the remote camera to carry out corresponding smooth movements. A similar example of processing the Wii Fit Balance Board data for the *WiiFitChair* model has also been presented in Chapter 6, with a detailed mathematical model.

**How to deal with the limitations of a particular interaction?**

It is possible that a selected natural interaction form has limitations or is not enough to provide the full functionalities for the entire design. This will lead to a "half-designed" system or a system with limited functions. For instance, the eye tracking system is only able to provide data with limitations, i.e. the gaze data on the screen is 2-dimensional so that can only be used to control camera pan and tilt functions. In this case, adding extra interaction form(s) to provide the rest functionalities of the design could be an effective solution, which is commonly referred as an multimodal design (e.g. adding head tracking to control camera zoom function for the gaze-driven model in Chapter 4 and for the *WiiFitChair* model in Chapter 6).

**What if the original setting is not available for conducting evaluations?**

Building and testing natural interfaces for complex industrial tasks is difficult. Some of this difficulty arises from industrial pressures which severely limit access to the usual operators of the equipment, as running user studies on real-world industrial settings will usually result in a long period of downtime for the equipment with a large amount of cost. Hence, development and testing must be done outside the actual setting. As introduced
in Chapter 4, developing a *Functional Physical Model* for the real-world setting can be an effective approach to conduct empirical user studies. This approach is to model the important features of the task as far as possible rather than attempting to model the task literally.

### 7.3 Future Work

In this thesis, we have demonstrated some initial success of using natural interaction to design camera viewpoint control models for teleoperation. A number of novel camera viewpoint control prototypes using different forms of natural interaction have been developed and evaluated. The results of the user studies have shown that using some forms of natural interaction (e.g. the *head motion control* in Chapter 3 and the *gaze-driven control* in Chapter 4) for camera viewpoint control can have significant improvements in performance for teleoperation tasks. Such models should not only be focused on designing teleoperation interfaces, as numerous possibilities can be explored on top of the initial success achieved in this work. For example, both head motion and gaze-driven control can also be used to develop interactive models for mobile and portable devices, such as smartphones. Moreover, investigating such natural interaction based control interfaces for multiple users to conduct collaboration in teleoperation is another very important research direction as we only addressed single-user issues regarding camera viewpoint control in teleoperation in this work. Furthermore, investigating other forms of interaction models also remains as a great potential for the future, such as gestures, touch and voice. Our attempts on the *WiiFitChair* and the foot-controlled model proposed in Chapter 6 have already demonstrated some preliminary steps to further this direction.

On the other hand, since the most successful camera viewpoint model introduced in this thesis is a combination of two different natural interaction forms (i.e. combining eye gaze and head motion as the natural interaction model introduced in Chapter 5), designing more multi-modal camera viewpoint control interfaces
would be an important direction with much potential value in this much bigger space. It is not necessary to just focus on combining different forms of natural interaction together, but to take advantage of these existing camera viewpoint models and combining them in the sense of using the most suitable model for a particular situation or in terms of the user preference. For example, an operator is able to use the gaze-driven model when they start working on the rock breaking task, but they can later switch to the autonomous tracking model (Chapter 5) when they feel tired of using gaze control or are performing more predictable tasks. Formal user studies would also be important to find the feasibility of these proposed models.

It is common in many teleoperation settings that multiple cameras are installed in order to cover the entire space of the remote environment, particularly in security and surveillance. Therefore, another future direction could be the investigation of natural interaction based viewpoint control models for multi-camera systems.

In the end, we hope the findings described in this thesis will encourage more research on such real-world teleoperation settings and may lead to the deployment of such natural interaction based camera control models in a practical context.
Appendix A

Experimental Material

The relevant material regarding the user experiments described in this thesis is attached in this section, including the consent form and the major questionnaire used in all the user studies.

The questions used in the questionnaire are derived from the IBM Computer Usability Satisfaction Questionnaire [86].
Natural Human Interaction for Remote Camera Control

Usability Study Consent Form

Introduction
This document concerns a usability study to be conducted at the School of Computer Science at the Australian National University. The study involves looking at the usability issues for an application, and its associated interaction techniques. This application involves the use of natural human interaction for controlling a Pan-Tilt-Zoom (PTZ) camera remotely.

What would be involved?
About 40 minutes of your time on one occasion will be needed. In this time you will get an introduction to the application and techniques mentioned above and also complete some tasks.

Data Collection and Contact Details
The main purpose of the usability study is to collect some data to enable useful information to be gained on the interface, the interaction techniques and tasks. We will give you a post-task questionnaire which may contain some questions of an identifying nature. You do not need to complete these or any of the other questions if you have any objections to them. After the entire experiment is completed, one year at most, data will be “deidentified”. Until that time, if you give your permission, your contact details will be retained for follow-up testing.

Data Use
The data collected will be used to draw conclusions about certain interaction techniques and the nature of the tasks. Any data collected, either raw or processed, may be used in a thesis and other publications. After “deidentification”, the participant will not be able to be identified from any data collected.

Risks
As the study is conducted in a carefully designed lab environment, all care will be taken to make participants as comfortable as possible, the nature of the interaction tasks. Some physical discomforts such as eye and muscle strain may occur with some people including, in rare cases, motion sickness. Participants are free to request that testing cease at any stage without explanation.

Your rights
You may ask for a copy of any data collected or research publications written. You may also end the test session or ask for a break at any time and request that any or all data collected be destroyed. You have the right to completely withdraw from the testing at any point. You can ask that your name be deleted from our contact list for future testing at any time.

This usability study has nothing to do with assessment and participation or non-participation will not directly affect your assessment in any course at ANU and it is completely voluntary. If you have any concerns with the ethics of this study please contact the ANU ethics committee by emailing Human.Ethics.Officer@anu.edu.au or calling 6125 2900.
Consent Section
By signing this form you agree that you have completely read both sides and understand and agree to it. You are also agreeing to participate in the study.

Please list any Special Considerations (eg any medical conditions you have which you would like to bring to the attention of the test supervisor)

Participant's Name: ________________________________

Signature: ________________________________ Date: __________
User Study Questionnaire

Thank you for taking the time to fill in this questionnaire. Please respond as truthfully as possible, as criticism is appreciated as much as positive feedback.

1 Personal Details

1. Name:

2. Age:

3. Sex: male female

4. Occupation (if you are a student, please specify your major):

5. How often do you use a computer?
   Never Occasionally Often (at least 2 hours a day)

6. How often do you play video games?
   Never Occasionally Often (at least 1 hour a day)

7. How much have you used a joystick or gamepad?
   Never Occasionally Often (at least 1 hour a day)

8. How much have you participated in teleoperation (remote control) tasks?
   Never Occasionally Often

9. How much have you used an eye-tracking based interface?
   Never Occasionally Often
2 Presence Questions

Used: "__________" for camera control.

1. It was natural to use this camera control method.

   1 2 3 4 5 6 7
   Strongly Disagree  Strongly Agree

2. I felt intuitive to use this camera control method.

   1 2 3 4 5 6 7
   Strongly Disagree  Strongly Agree

3. It was easy to learn to use this camera control method.

   1 2 3 4 5 6 7
   Strongly Disagree  Strongly Agree

4. It didn’t take long to get used to this camera control method.

   1 2 3 4 5 6 7
   Strongly Disagree  Strongly Agree

5. It was simple to find the stuff I wanted to see in the remote place using this camera control method.

   1 2 3 4 5 6 7
   Strongly Disagree  Strongly Agree

6. I was able to gain enough visual information from the video stream using this camera control method.

   1 2 3 4 5 6 7
   Strongly Disagree  Strongly Agree

7. I was able to obtain enough situational awareness for the teleoperation task using this camera control method.

   1 2 3 4 5 6 7
   Strongly Disagree  Strongly Agree
8. I didn’t have to pay much attention or consciousness on the camera control using this method when conducting the other control task by hands.

1  2  3  4  5  6  7  
Strongly Disagree  Strongly Agree

9. It was not distracting to use this camera control when conducting the other control task by hands.

1  2  3  4  5  6  7  
Strongly Disagree  Strongly Agree

10. I felt the zooming function of this camera control was useful.

1  2  3  4  5  6  7  
Strongly Disagree  Strongly Agree

11. I could effectively complete the rock pushing task using this camera control method.

1  2  3  4  5  6  7  
Strongly Disagree  Strongly Agree

12. I was able to quickly complete the rock pushing task using this camera control method.

1  2  3  4  5  6  7  
Strongly Disagree  Strongly Agree

13. I was able to efficiently complete the rock pushing task using this camera control method.

1  2  3  4  5  6  7  
Strongly Disagree  Strongly Agree

14. I believe I could become productive using this camera control method.

1  2  3  4  5  6  7  
Strongly Disagree  Strongly Agree

15. Overall, I am satisfied with this camera control method.

1  2  3  4  5  6  7  
Strongly Disagree  Strongly Agree
3 Open Ended Questions

1. Please rank the camera control methods you used in the experiment based on your experience (the best one goes first).

2. In what way do you feel those control methods either enhanced, or detracted the performance of the tasks?

3. Do you think those control methods could be improved, if so, how?

4. Do you have any other comments about the control methods or anything related to them?
References


http://www.so-touch.com/


REFERENCES


REFERENCES


REFERENCES


   http://depthjs.media.mit.edu/

   http://www.merl.com/projects/DiamondTouch/

   http://actu.epfl.ch/news/connecting-kinects-for-group-surveillance/

   http://www.digiwall.se/

International Conference on Field and Service Robots (FSR 2009), pp. 1-10.

Multiple Home Appliances. Proceedings of 8th TC.13 IFIP International
Conference on Human-Computer Interaction (INTERACT 2001), pp. 727-
728.

[38] Endert, A., Fiaux, P., Chung, H., Stewart, M., Andrews, C. and North, C.,
2011. ChairMouse: Leveraging Natural Chair Rotation for Cursor Naviga-
tion on Large, High-Resolution Displays. Proceedings of 29th ACM SIGCHI
Conference Extended Abstracts on Human Factors in Computing Systems

   http://www.sr-research.com/EL_II.html

[40] EyeToy, Playstation, Sony, 2011.
   http://uk.playstation.com/ps2/peripherals/detail/item51697/EyeToy-
   USB-Camera/
REFERENCES


164


REFERENCES


http://www.apple.com/ipad/


http://www.apple.com/iphone/


REFERENCES


[77] Kinect, Microsoft, 2011.

http://www.xbox.com/kinect/


REFERENCES


http://www.microsoft.com/surface/


http://nuigroup.com/log/comments/forums_launched/.


    http://www.naturalpoint.com/

    http://wiki.nuigroup.com/Natural_User_Interface/


REFERENCES


http://www.pelco.com/

http://www.sensible.com/haptic-phantom-omni.htm/

http://www.sensible.com/phantom-premium-1-5.htm/


REFERENCES


http://www.tobii.com


http://www.nintendo.com/wii
http://www.wiifit.com/

http://www.wiiuse.net/


REFERENCES


178