symmetry

MDPI

Article

Symmetric Evaluation of Multimodal Human–Robot Interaction with Gaze and Standard Control

Ethan R. Jones [1], Winyu Chinthammit [1,*] , Weidong Huang [2,*], Ulrich Engelke [3] and Christopher Lueg [1]

[1] School of Technology, Environments and Design, University of Tasmania, Hobart, TAS 7005, Australia; ethan.jones@unsw.edu.au (E.R.J.); christopher.lueg@utas.edu.au (C.L.)

[2] School of Software and Electrical Engineering, Swinburne University of Technology, Hawthorn, VIC 3122, Australia

[3] CSIRO, Kensington, WA 6020, Australia; Ulrich.Engelke@data61.csiro.au

* Correspondence: winyu.chinthammit@utas.edu.au (W.C.); weidonghuang@swin.edu.au (W.H.)

Received: 16 October 2018; Accepted: 28 November 2018; Published: 1 December 2018

check for updates

Abstract: Control of robot arms is often required in engineering and can be performed by using different methods. This study examined and symmetrically compared the use of a controller, eye gaze tracker and a combination thereof in a multimodal setup for control of a robot arm. Tasks of different complexities were defined and twenty participants completed an experiment using these interaction modalities to solve the tasks. More specifically, there were three tasks: the first was to navigate a chess piece from a square to another pre-specified square; the second was the same as the first task, but required more moves to complete; and the third task was to move multiple pieces to reach a solution to a pre-defined arrangement of the pieces. Further, while gaze control has the potential to be more intuitive than a hand controller, it suffers from limitations with regard to spatial accuracy and target selection. The multimodal setup aimed to mitigate the weaknesses of the eye gaze tracker, creating a superior system without simply relying on the controller. The experiment shows that the multimodal setup improves performance over the eye gaze tracker alone ($p < 0.05$) and was competitive with the controller only setup, although did not outperform it ($p > 0.05$).

Keywords: multimodal interaction; eye tracking; empirical evaluation; human–robot interaction

1. Introduction

Multimodal interaction for effective human–robot interaction (HRI) is a field with considerable potential. It has the potential to offer intuitive interaction through taking advantage of several interaction devices acting in support of each other. However, it has not reached the point of being generally deployed and a significant line of research is needed to fully understand the usefulness of multimodal techniques. This research attempted to contribute to this field by investigating multimodal interaction with two different techniques, hand and gaze control.

Eye gaze based systems have been examined for interface control [1] and evaluation [2,3] since the 1980s, shortly after computer technology advanced to the point of making the creation of such systems possible. Gaze-based interaction has many applications in a range of fields such as engineering and human–computer interaction and offers many opportunities as well as challenges for research [4,5]. Users will generally look at what they wish to interact with, and even reliably do so before the target is reached using conventional mouse movement [6]. Gaze based interaction could theoretically be faster than mouse interaction at selection tasks if accuracy and confirmation issues were resolved.

Research into eye movement has shown that eye movement through a scene is related to scene content. The saliency map locates important scene sections based on their "conspicuity" (relative number of unusual features) [7]. These maps are somewhat predictive of actual eye motion through a scene, while still falling short enough that it seems likely there is more at play [8]. At least one of these other factors seems to be the task, with research showing that eye motion varies significantly with task [9,10]. The task of an individual viewing a scene can even be predicted through their eye motion [11]. Given that it seems that task and gaze patterns are naturally connected it can be said with a degree of certainty that gaze based interaction has the potential to be highly intuitive.

Gaze based interaction faces two key issues: spatial accuracy and selection. The speed with which gaze interfaces can reach a target is often offset by its inaccuracy [12–14]. Gaze targeting cannot achieve the precision of mouse selection due to limitations on how precisely the gaze target can be determined by the hardware. In addition, when a gaze interface is used in a real environment, a number of factors, such as lighting and calibration quality, can introduce further inaccuracies. Ultimately, this generally results in a need for some form of target disambiguation [15] and correspondingly a significant reduction in interaction speed.

Issues with selection in gaze interaction centre around what is referred to as the "Midas touch" problem [16,17]. A user will often fixate on something without intending to interact with it. If a gaze-based interaction system is set up to trigger interactions when it detects a fixation on that point, these fixations will inadvertently result in the system taking some kind of action. This results in a situation where the user must uncomfortably avert their gaze in order to avoid directly looking at interface elements, instead attempting to view the interface through their peripheral vision.

In this paper, a method of mitigating these two issues with multimodal interaction is presented and assessed. The multimodal setup was designed such that it attempted to mitigate the weaknesses of gaze tracking without compromising its strengths. The focus was thus on the eye gaze tracker, using the controller in an auxiliary role to control selection and manage some of the accuracy issues in gaze tracking. In the evaluation, the multimodal setup was compared to the eye gaze tracker to see whether the multimodal setup is actually capable of mitigating the issues mentioned in single modality eye tracking. The controller alone was also used as a base control condition to see whether the multimodal setup can compete with a standard setup.

The remainder of this article is organised into the following sections: Section 2 discusses some past eye gaze control and multimodal interaction research, as well as where more work is needed. Section 3 covers the details of the experiment performed in this research. Section 4 reports the results of this study and presents analysis of those results. Section 4 discusses the implications of these results. Section 6 summarises the final conclusions derived from this study.

2. Background

Much research has been done to review latest developments and recent advancements in gaze-based interactions. Duchowski [4] categorized gaze-based interaction into four forms: diagnostic; active, such as selection and look to shoot; passive, such as gaze-contingent displays; and expressive, such as gaze synthesis. It was found that latest diagnostic and interactive possibilities provided by gaze-based interaction has opened an additional bidirectional channel for the user to receive rich information such as that associated with cognition and expression. Asan and Yang [2] reviewed research that used eye trackers to evaluate usability of Health Information Technology. They found that eye tracking could provide valuable data for evaluation. However, there is a huge gap in using it to evaluate clinical and mobile health systems and applications in natural settings. More than a decade ago, Morimoto and Mimica [18] reviewed the state of gaze interaction technology with respect to practical use by the average user. They described how current systems were limited by the need for constrained head movement and constant recalibration, and accordingly gaze tracking was not yet capable of delivering a high enough quality of experience for general use, despite its apparent potential. While gaze interaction technology has advanced over the following years, it remains niche, evidently lacking practicality for general use. However, there has been substantial research into resolving its limitations.

Tuisku et al. [19] examined the use of a somewhat multimodal setup that made use of facial movements for selections. This system found some success as a method for hands off Human–Computer Interaction (HCI). Stellmach and Dachselt [15] compared a combined gaze and touch pad system with a head and gaze only system using a movable magnifier. Both systems were designed to improve accuracy, with the touch pad system achieving better performance. Despite this the head based system was perceived by users as having better performance, following a trend in gaze research for users to disproportionately favour fully hands off systems. Velichkovsky et al. [17] attempted to differentiate between focal fixations and other fixations by examining the user's eye saccades, achieving some success in dealing with the "Midas touch" issue.

Two of the three studies mentioned studies take advantage of multimodal interaction to tackle the weaknesses of standard gaze interaction. Whilst they have their origins in combined voice and gesture interaction, as created by Bolt [20], multimodal interface research has come to examine a huge set of interface combinations. Turk [21] discussed the progress of multimodal interaction research and outlines some of the ways in which it has been used. Despite this broad usage, Turk stated that "Multimodal integration methods and architectures need to explore a wider range of methods and modality combinations", an issue that this study aimed to address by examining the combination of eye gaze and controller modalities.

One of the possible areas in which multimodal interaction may gain benefit is that of cognitive load. The concept of cognitive load originated in the field of education, as a way of modelling the way people learn to construct better teching methods [22–25]. In this model, humans have a working memory, which can only handle a small set of interacting elements, and long-term memory, which stores a large number of schemas that can be used to reduce several interacting elements into a single element in working memory.

It is thought that, by distributing their communication through multiple pathways, a user might reduce their overall cognitive load with multimodal interaction [26]. Effectively, a schema models the complex multimodal interaction in one element, rather than requiring a multiple step single modality interaction that would require several elements to be kept in storage. The study mentioned gave participants tasks of increasing complexity and measured the proportion of their interactions that were multimodal. The system made use of speech and pen input modalities to make up the multimodal system. It was found that the proportion of multimodal interactions did increase with task complexity, indicating that was the easier method to use to relay complex information. Therefore, in our study, we also examined how cognitive load was affected by different modality conditions.

3. Experiment Methodology

We performed an experiment to determine the relative performance of single versus multimodal control of a robot arm. In this section, we discuss in detail the experiment methodology.

3.1. Design Overview

Three interface setups were used to compare three modality conditions: two that used each of the individual devices, controller and eye tracker alone, and one that used both devices. These interaction modalities are in the following referred to as *controller*, *gaze*, and *multimodal*, respectively.

Three different task complexities were used to analyse how these interfaces scaled with task complexity, which are referred to as *simple*, *moderate*, and *complex*. We had one task for each complexity level. We employed a within-subject design, which means that each participant performed 3 (modality) × 3 (task complexity) = 9 experiment tasks to fill in a full factorial design, between interface device and task complexity. Before performing any tasks, the participant was given a chance to practice with the device configuration in a practice task, for which no data were recorded.

All tasks were completed before moving on to the next interaction modality. Tasks were assigned to participants in a counter-balanced fashion. In addition, the order in which the devices were used was varied according to a Latin Square design to control for learning the task with the earlier interface devices [27]. During the experiment, task response times and cognitive load were measured to test the effectiveness of the interfaces. More specifically, cognitive load was measured using the NASA TLX [28,29]. We also distributed a questionnaire asking participants for demographic information and their preferences of the interfaces.

3.2. Participants

Twenty participants (six female) were recruited on a completely volunteering basis from the ICT student population of the University of Tasmania. At the time of participation, participant aged ranged between 18 and 42 with a mean age of 26.25. On the self-rated experience scale from the survey, the average controller experience was 3.1, with at least one participant rating at each level. Experience with gaze was rated 1.15, with three ratings of 2 and no higher ratings.

3.3. Apparatus

This study made use of The Eye Tribe eye tracker (https://theeyetribe.com/). It has an accuracy of 0.5–1.0 degrees and a recording frequency of 60 Hz. This device was used alongside the Arbotix Commander 2.0 controller (http://www.trossenrobotics.com/p/arbotix-commander-gamepad-v2.aspx) to control the Arduino WidowX Robot Arm (http://www.trossenrobotics.com/widowxrobotarm). The controller is the standard controller designed for use with the WidowX. Video was recorded with a Microsoft LifeCam Cinema webcam. All these devices were integrated through a MATLAB interface.

All tasks were performed on a 6 by 5 "chess board" grid, with the tasks following a chess analogy. The user interface was displayed on a computer in the form of a live video of the robot arm and the board, annotated with relevant information. Specifically squares could be highlighted around the edges with any colour and a message could be displayed. The use of the highlights varied by task and interface; however, the messages always acted to indicate when the system was waiting for or had received input. This feedback message was in addition to the feedback of the robot arm actually carrying out the task. Note that, during the task, based on the input information received from different modalities, the system would send signals to move the robot arm automatically. It had pre-programmed positions for each square on the chess board, which was always placed in the same relative location.

3.4. Tasks

The three tasks are summarized in Table 1 and are discussed in more detail in this section. We created three different tasks that we loosely classify as *simple*, *moderate*, and *complex*.

Table 1. Summary of the three tasks. Limited selection means the participant could only move the cursor on to squares that the piece could move to. Multiple pieces means the participant was asked to manipulate multiple pieces to complete the task. Minimum moves is the fewest moves in which it was possible to complete the task.

Task Complexity	Limited Selection	Multiple Pieces	Minimum Moves
Simple	Yes	No	3
Moderate	No	No	4
Complex	No	Yes	Varies (4 to 6)

3.4.1. Simple Task

The participant was asked to navigate a "knight" chess piece from a certain starting square to another square using the least possible legal chess moves. The goal square that needed to be reached was marked with a green outline. All the squares that the piece could reach from its current position (the valid squares/moves) were marked with a yellow outline. The start and end positions for the three simple task setups are shown in Table 2. From the camera these coordinates are such that the x-axis is away from the camera and coordinates begin in the bottom left. All of these setups require an exact minimum of three moves to complete.

The design of this task attempted to present the participant with an easy task that still required at least some cognitive load. The interaction setup, with its restriction to places that are valid moves for a knight piece, also acted as an introduction to knight movement for those participants without prior chess knowledge. The multimodal setup for this task was also simple, corresponding to the simple task, with both modalities allowing for square selection (though only the controller could perform confirmation). A screenshot of the Graphic User Interface (GUI) during the simple task is shown in Figure 1.

Table 2. Setups for simple and moderate tasks, given as board coordinates for the start and target positions.

Simple Task			Moderate Task		
	Start	**Target**		**Start**	**Target**
Task 1	(2, 2)	(5, 2)	Task 1	(5, 0)	(4, 1)
Task 2	(0, 0)	(5, 4)	Task 2	(0, 0)	(2, 2)
Task 3	(3, 4)	(0, 2)	Task 3	(1, 3)	(0, 4)

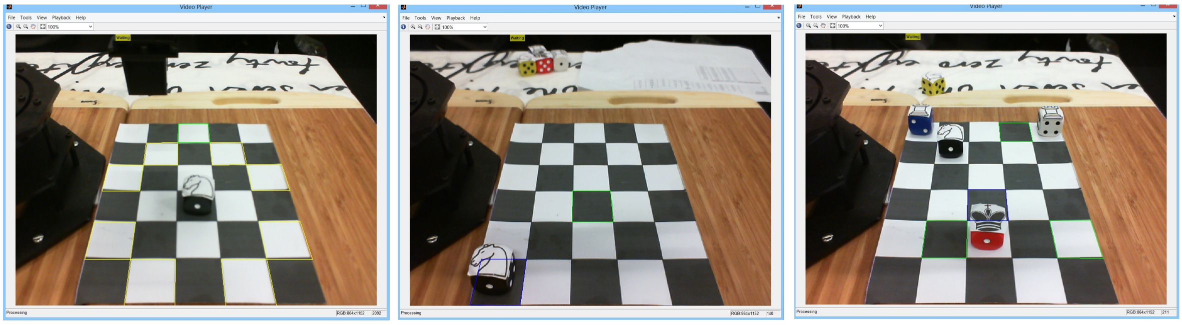

Figure 1. GUI examples at the beginning of: the simple task (**left**); the moderate task (**middle**); and the complex tasks (**right**) for the controller alone.

3.4.2. Moderate Task

As with the simple task, the participant was asked to navigate a knight chess piece from one square to another with the fewest chess moves they could manage. However, in this case, less of the interaction was automatic, simulating a system with less prior knowledge of the task, therefore requiring more information from the user. The initial and destination squares for the setups also contributed to increased complexity, with a minimum of four moves being needed to complete the task. Again, the target square was marked with a green outline. The start and end positions for the three moderate task setups are shown in Table 2.

The moderate task was designed to be only a small step up from the simple task whilst being clearly the more complex of the two. It allowed selection of any square, thereby requiring participants to determine for themselves where the knight may move, increasing cognitive load and the range of interaction options. The specific task setups also required a higher minimum number of moves, such that the task should be more complex than the simple task even if the control scheme was the same. The multimodal setup for this and the complex task more clearly delineates the roles of the two devices and does not give much choice to the participant as to which modality should be used for a task. This design choice is rooted in the goal of creating a setup that mitigates gaze weaknesses rather than replacing gaze with the controller. If the participant were allowed to use the controller for the full task some participants may do so, skewing the comparison and creating ambiguity as to whether gaze weaknesses were really mitigated. A screenshot of the GUI during the moderate task is shown in Figure 1.

3.4.3. Complex Task

Unlike the other two tasks, for the complex task, the participant was asked to move multiple pieces to reach a solution. These pieces had to be arranged to achieve a "checkmate"-like arrangement against a king piece. The target king piece could not move or take the participant's pieces, but the participant was not allowed to take it either. Instead, they had to arrange their pieces such that both the square the king was on and the squares surrounding it (including diagonal) could be reached by at least one of their pieces in a single move, or already contained one of their pieces. A hint was added by outlining in green a set of squares that, should the correct pieces be placed in them, would complete the task. The currently selected piece was marked with a cyan outline. The setups for the complex task are shown in Table 3.

Table 3. Complex task setups, given as board coordinates for the pieces and hints.

	Participant Pieces	King	Hints	Minimum Moves
Task 1	Knight (4, 1), Rook (5, 0), Rook (5, 4).	(1, 2)	(1, 1), (1, 4), (5, 3).	5
Task 2	Knight (0, 4), Rook (5, 0).	(0, 2)	(1, 2), (2, 2).	6
Task 3	Knight (5, 0), Knight (2, 3), Knight (5, 3).	(0, 0)	(3, 0), (2, 2), (1, 2).	4

The complex task was designed as a significant step up from the other two tasks, in order to make it more likely that some effects from the increased complexity would be seen. It was, however, still restricted by the limitation that the experiment was to be run in a single session and each task needed to be completed three times. It therefore had to be restricted to a level of complexity such that most participants could complete the task within a few minutes. The purpose of the hint squares was to keep participant time spent on this task relatively low. A screenshot of the GUI during the complex task is shown in Figure 1.

3.5. Interaction Modalities

As each interaction modality was set up according to the task complexity, we give details in this section on how different interaction modalities were set up, along with flow charts for participant interaction, as shown in Figures 2–4.

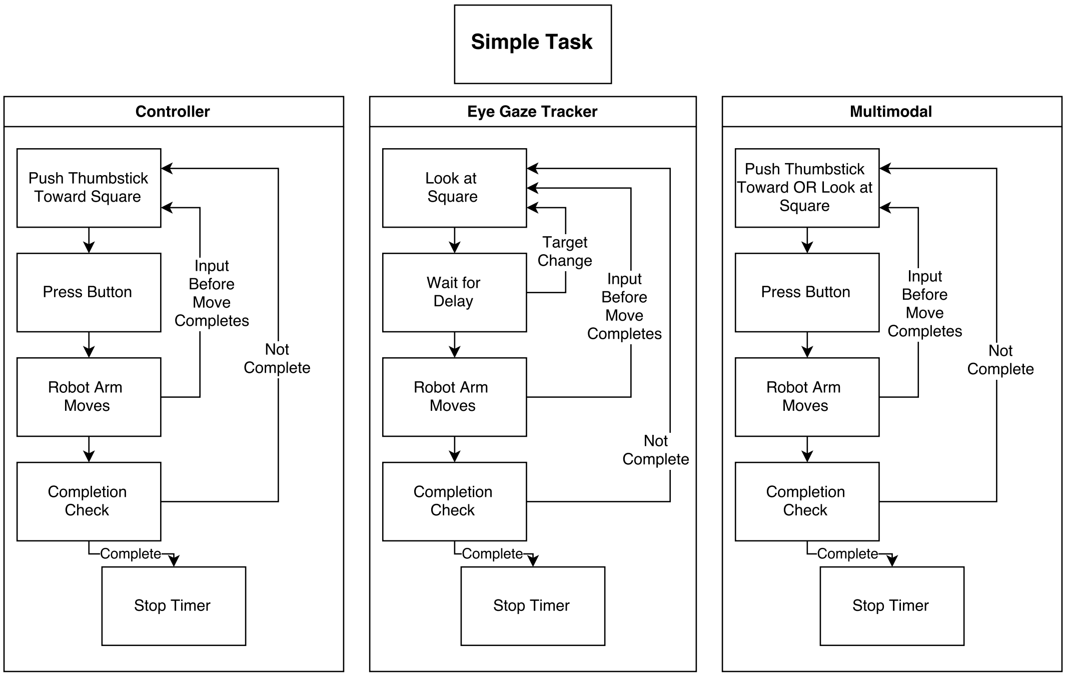

Figure 2. A flow chart of participant interaction in the simple task.

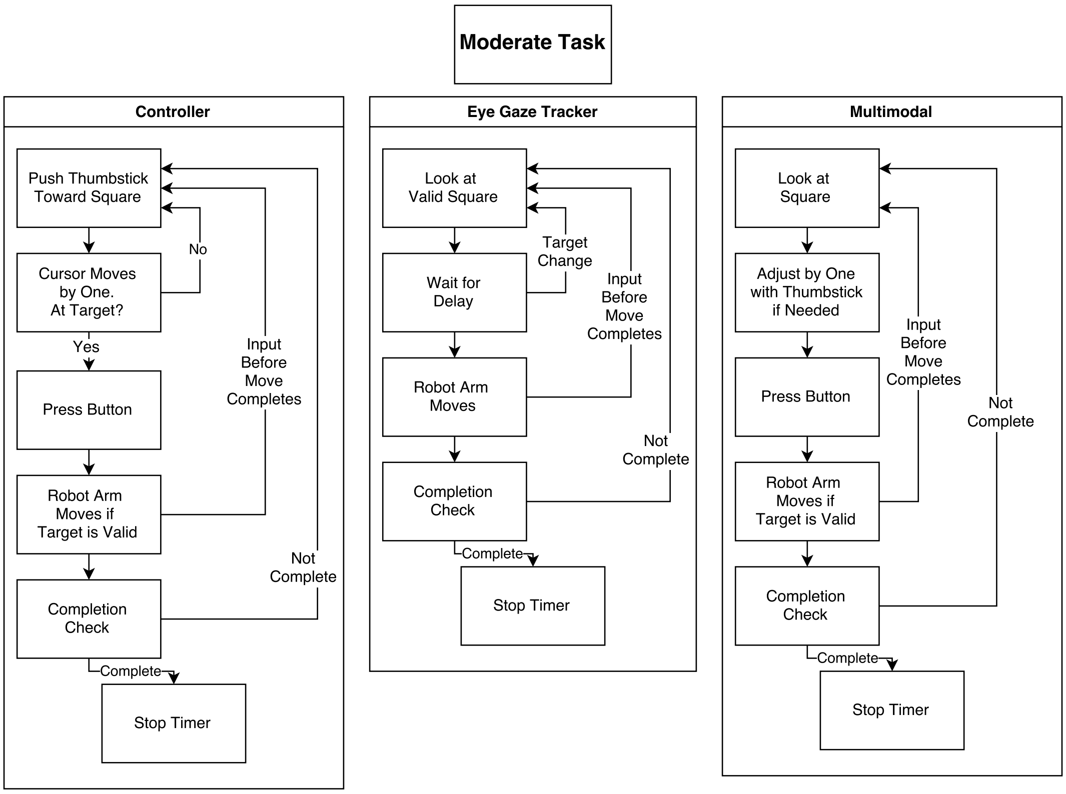

Figure 3. A flow chart of participant interaction in the moderate task.

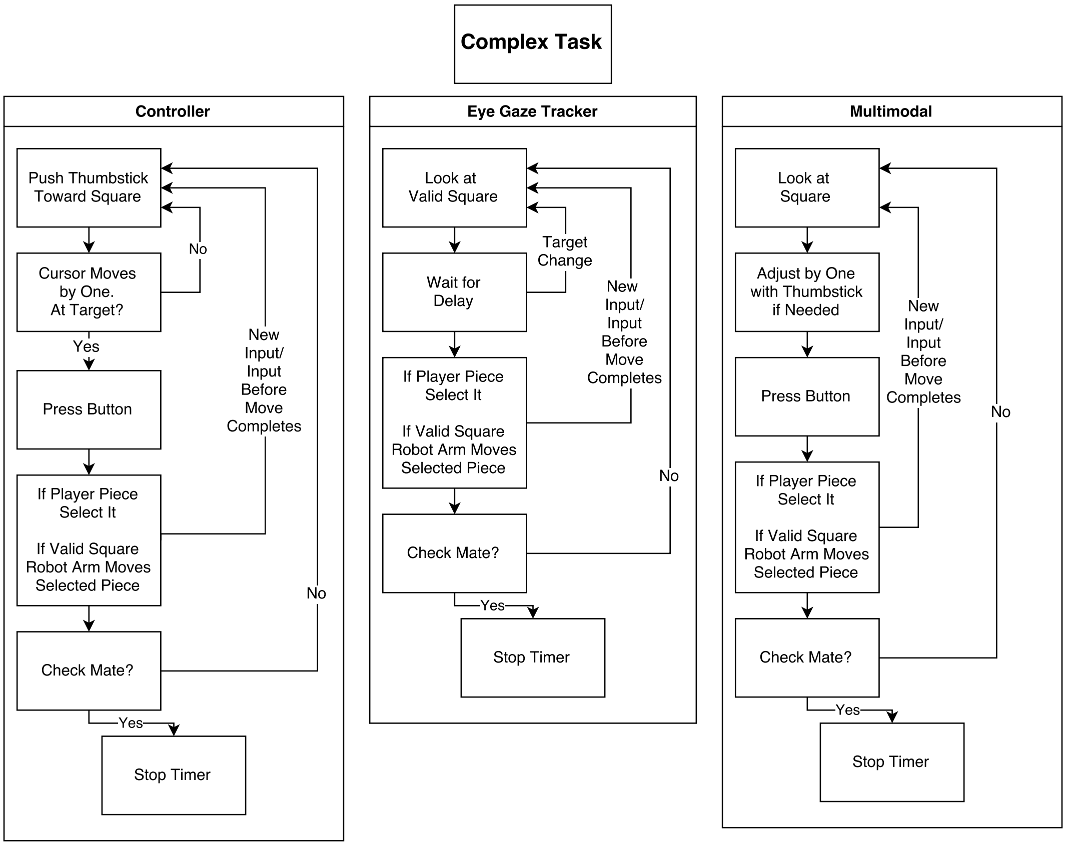
Complex Task
Controller
Push Thumbstick Toward Square
Cursor Moves by One. At Target?
No
Yes
Press Button
New Input/ Input Before Move Completes
If Player Piece Select It
If Valid Square Robot Arm Moves Selected Piece
No
Check Mate?
Yes
Stop Timer
Eye Gaze Tracker
Look at Valid Square
Wait for Delay
Target Change
New Input/ Input Before Move Completes
If Player Piece Select It
If Valid Square Robot Arm Moves Selected Piece
No
Check Mate?
Yes
Stop Timer
Multimodal
Look at Square
Adjust by One with Thumbstick if Needed
New Input/ Input Before Move Completes
No
Press Button
If Player Piece Select It
If Valid Square Robot Arm Moves Selected Piece
Check Mate?
Yes
Stop Timer

Figure 4. A flow chart of participant interaction in the complex task.

3.5.1. Controller

When using the controller alone, there were two main concepts used in the control schemes. The first was used in the simple task. If the thumbstick was in its deadzone (not being pushed), there were no farther annotations and pressing the selection button had no effect. However, if the thumbstick was pushed in any direction, the knight move square closest to that direction would be outlined in blue. If the square was also a valid move then pressing the selection button would cause the piece to be moved to that square. This is outlined in Figure 2.

For the moderate and complex tasks, the basis for the control scheme differed. A square that acted somewhat similar to a cursor was outlined blue. This cursor square could be navigated one square at a time to adjacent squares by pushing the control stick in the direction of the square desired. A 500 ms delay timer prevented the square from moving more than one place when the thumbstick was pushed, the delay being reset if the control stick was returned to the deadzone. When the cursor square was a valid move for the selected piece and the selection button was pressed the piece would be moved to the target square. This is outlined in Figure 3. In the complex task, pressing the selection button over an unselected piece would change the selected piece to be that piece. This is outlined in Figure 4.

3.5.2. Gaze

As with the controller alone, there were two main interface concepts used with the gaze tracker alone, one for the simple task and one for moderate and complex. In the simple task whenever the eye gaze tracker had a clear enough image to give a gaze target pixel that pixel was used to select a target square. This square was the valid square closest to the pixel that was targeted by the participant's gaze. This square was marked with a red outline. Whenever the eye gaze tracker registered a fixation the current target square would be selected, causing the piece to be moved to that square. This is outlined in Figure 2.

When performing the moderate or complex tasks the target could be moved more freely. Again, whenever there was a clear gaze target pixel, that pixel was used to select a target square. In these tasks this square was whichever square on the chess grid was closest to the pixel, regardless of whether it was a valid square. As before, when a fixation was detected the square would be selected, however this would only have an effect if the current target square was a valid square or contained an unselected piece. In these cases, the selected piece would be moved or the selected piece changed, respectively. This is outlined in Figures 3 and 4.

3.5.3. Multimodal

As with the other interface schemes the multimodal setup differed considerably between the simple task and the moderate and complex tasks. In the case of the simple task, a square could be outlined red in the same way as with gaze alone or blue in the same was as with the controller alone. When the selection button on the controller was pressed, a selection would be made if either of these highlights existed and the piece moved. The controller's blue square took priority in this, with the piece moved to it if it existed. The piece was only moved to the red gaze square if the selection button was pressed and the controller's thumbstick was in its deadzone. This interface effectively gave the participant the choice of which device to use for the task. This is outlined in Figure 2. The device used for each selection was recorded.

The interface for the moderate and complex tasks did not allow such choice, instead combining the two devices to complete the task. A square would be outlined red as it was when using the eye gaze tracker alone, however fixations would not trigger selections. Instead, the controller's selection button was used to make selections. In addition, if the controller's thumbstick was not in its deadzone, the target square would be displaced in the direction the thumbstick was pushed. This resulted in the target square at any given time being the one resulting from the gaze target shifted up to one square vertically and one square horizontally based on the direction on the thumbstick. The results of selections were the same as for the other two devices, moving the selected piece if a valid square was targeted or changing the selected piece if another of the participant's pieces was targeted. This is outlined in Figures 3 and 4.

3.6. Procedure

The experiment was conducted individually in a dedicated and controlled laboratory environment. First, the participant read the information sheet and signed the consent form. Then, an introduction about the experiment, tasks, interfaces and procedure was given. After that, the participant was given time to ask questions and practice sample tasks with the interfaces to familiarise themselves with the devices he/she would be using through the experiment. The practice task was designed to allow the user to learn the interaction mechanism without actually relating to the real tasks to be performed beyond that.

Once ready, the participant indicated to the experimenter to start the experiment. All experiment sessions were performed by a single researcher who was present through the entire process. The researcher made use of a prepared script to deliver instructions to the participant throughout the experiment. All the participants were asked to perform the tasks as quickly as possible without sacrificing the task quality. They were told that their task response times were recorded. After each task was completed for each modality interface, the NASA TLX data were collected. After all tasks were completed, a questionnaire was distributed. The questionnaire included questions for demographic information such as age and gender and user preferences. More specifically, participants were asked for five-point scale ratings of their previous experience with controller and gaze input with 1 being the least experienced. Participants were also asked to rank the three interface setups from 1 to 3 for ease of use and for task effectiveness with 1 being the best. Finally, there was space for comments if participants wished to add any. The results of this questionnaire were considered secondary data to be compared to the primary completion time and NASA TLX data.

4. Results

We analysed the experiment outcomes and specifically the completion time, mental effort in terms of NASA TLX scores, and the questionnaire results. Analysis of variance (ANOVA) with repeated measures was employed to test effects of modality in each of the task levels at the significance level of 0.05. The error bars in all figures indicate 95% confidence intervals.

4.1. Completion Time and Cognitive Load

The mean values for completion time of all nine tasks are shown in Figure 5. It can be clearly seen that the controller has the lowest times and the gaze has the highest. The complex task also took participants longer than the other tasks. Statistical tests showed that there were no significant differences between conditions for both the simple and moderate tasks ($p > 0.05$). However, differences for complex tasks were statistically significant ($p < 0.05$).

The mean cognitive load scores are presented in Figure 6. There is a clear increase in values between the controller, multimodal and gaze, with the latter getting the highest workload ratings. In contrast, the workload scores for the complex task are not as much higher than those of the other tasks as the time score is. Statistical tests showed that there were significant differences between conditions for all tasks ($p < 0.05$).

In summary, the controller has outperformed the other devices, and the multimodal setup has outperformed gaze alone. However, it has not done so significantly for completion time. In terms of workload ratings, both the controller and multimodal systems outperformed the gaze alone by a statistically significant amount ($p < 0.05$). The only statistically significant result in terms of completion time, however, is that the controller outperformed gaze on the complex task. In no instance did the controller significantly outperform the multimodal system, in terms of completion time or workload.

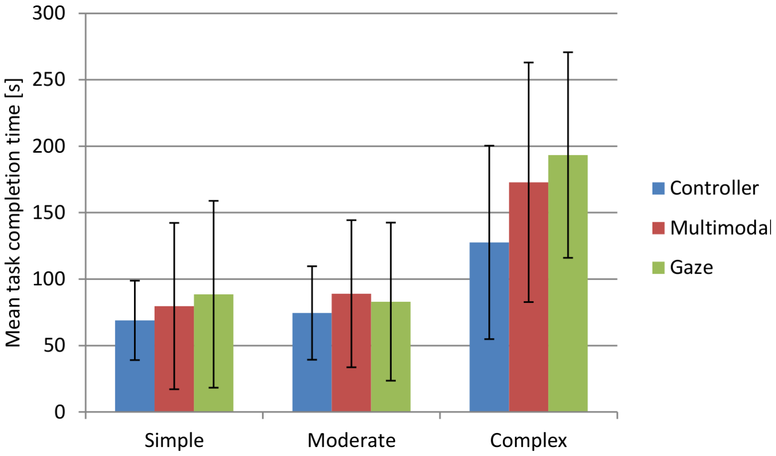

Figure 5. Mean times of modalities for all task complexities.

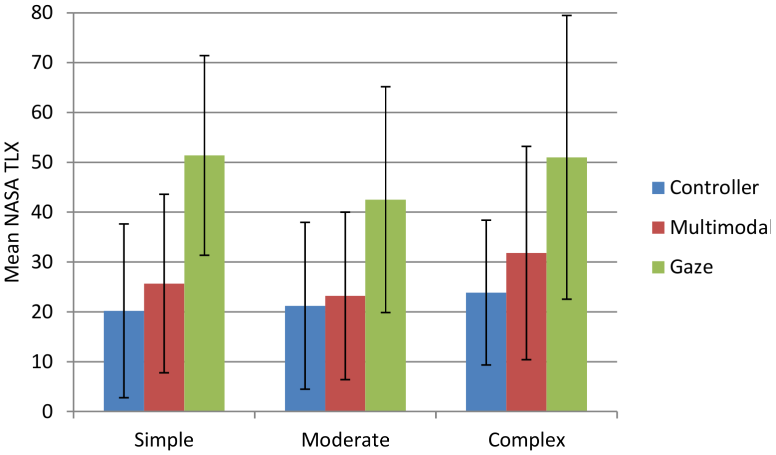

Figure 6. Mean NASA TLX scores for all task complexities and interaction modalities. Scores can range from 0 to 100.

4.2. Questionnaire Results

Participants were asked to rank the interaction modalities for each task in terms of *ease of use* and *task effectiveness*. The mean values for the various tasks and complexities in both ease of use and effectiveness rankings are shown in Figures 7 and 8, respectively.

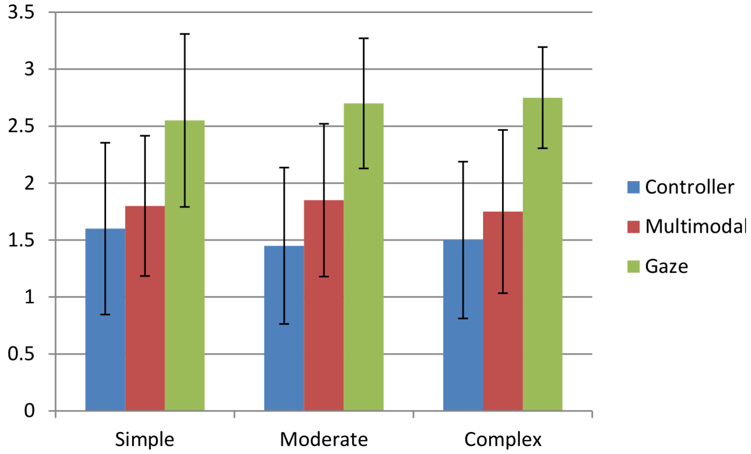

Figure 7. Mean *ease of use* rankings for all interaction modalities and tasks.

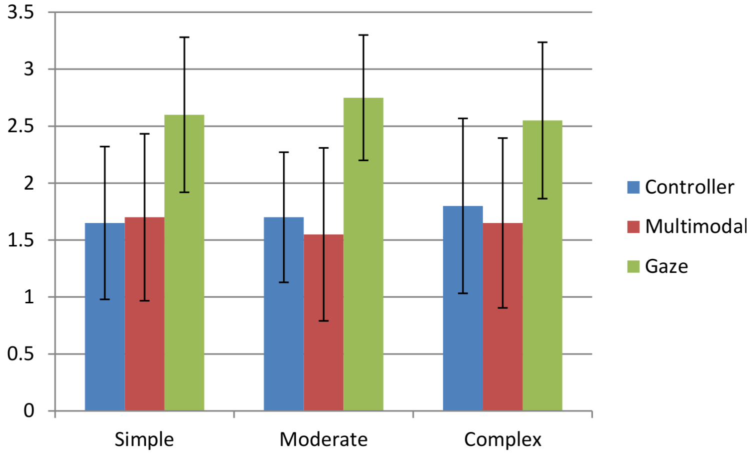

Figure 8. Mean *effectiveness* rankings for all interaction modalities and tasks.

Once again, overall gaze alone scored poorly compared to the other control modalities, although, interestingly, some participants ranked it as the best. While the controller achieved better scores in ease of use ranking and effectiveness ranking for the simple task, multimodal achieved the best score of effectiveness rankings for the moderate and complex tasks. The participants' preferences here differ from what completion time, cognitive load and even ease of use ranking data would indicate should be expected.

ANOVA analysis of the ease of use and effectiveness rankings indicated that both the controller and multimodal modalities achieved statistically significantly better results than the eye gaze tracker both in ease of use and effectiveness rankings for all tasks ($p < 0.05$). Between the controller and multimodal setups, no results were statistically significant ($p > 0.05$).

5. Discussion

5.1. Quantitative Data

In terms of cognitive load, we observed that multimodal interaction succeeded in outperforming gaze alone. The subjective rankings also support this observation and completion times were in favour of the multimodal interface, although not by a statistically significant margin. The gaze only and multimodal interfaces were both limited by gaze tracker inaccuracies. The degree to which this was an issue varied greatly by user. In the case of the gaze only setup, spatial inaccuracy worsened the "Midas touch" issue, as inaccurate detection of gaze target resulted in the robot arm moving toward the wrong target. With the multimodal interface button confirmation mitigated the "Midas touch" issue. Without the "Midas touch" issue, users were usually able to compensate for gaze tracker inaccuracy fairly easily by looking slightly off target (although the need to do this was far from ideal). Therefore, the multimodal interface has acted as a significant improvement on the gaze alone and, given that it was primarily gaze based, has successfully mitigated the weaknesses of the eye gaze tracker: spatial inaccuracy and the "Midas touch" problem. Clearly multimodal interaction is potentially useful as a method of mitigating weaknesses in gaze interaction at least. It would be worth investigating other interaction types with such extreme strengths and weaknesses as gaze to see if a multimodal system could achieve significant improvement in them as well.

Our study showed that the multimodal interface was not able to outperform the controller only interface, as the controller and multimodal modalities did not differ by a statistically significant margin. However, to a reasonable extent, this is in itself an interesting result, as it means the multimodal system achieved a comparable result to the controller alone. While, of course, the multimodal system does make use of the controller in this study, the controller was placed in an auxiliary role. It confirms selections and allows for some fine adjustment to the target, but cannot be used for overall control.

However, these small additions combined with gaze control did achieve a result that had significantly lower workload ratings than gaze alone.

In terms of completion time, there is little indication of which interaction modality was faster for the simple and moderate tasks. However, the controller was statistically faster than the other two for the complex task. In addition, the controller and multimodal modalities scored closely in both ease of use and effectiveness rankings. What makes this result interesting is the fact that participants had far higher average experience with controllers than eye gaze trackers, as assessed by the questionnaire. The controller is, as one of the standard interaction devices used today and the standard device for controlling the robot arm, a very tough competitor. Design of controllers has been refined; such devices are very stable and many participants use them regularly in everyday life. The eye gaze tracker is a new and less precise device that 17 of the 20 participants had not encountered until the experiment.

While in our study the controller was not significantly outperformed by the multimodal setup, systems that can outperform standard devices are likely worth further investigation. Hoste et al. [30] compared a multimodal speech and gesture interface to a conventional controller setup for typing. Their results, despite using a very different multimodal setup, echo those here in that the multimodal setup performed very similarly to the controller alone. They also found participants felt the multimodal setup was "most efficient", similar to the results for *effectiveness* rankings in our study. On the one hand, there was still room in our study to further refine the multimodal interaction setup, such as use a better eye tracker, which may push it ahead of the controller. On the other hand, participants could have been given more time to properly familiarise themselves with gaze based control. Given that the participants had greater average past experience with controllers, the multimodal setup would have had a greater learning curve than the controller and could have been disadvantaged in such an experiment. Alternatively, a study involving more participants with greater eye gaze tracker experience could be used to examine this possibility.

5.2. Qualitative User Feedback

A large component to the ratings and completion time for gaze and, to a lesser extent, multimodal setups was quality of calibration. Some participants were able to get very high quality calibrations such that the eye gaze tracker could reliably distinguish which square they were looking at. This effectively meant the tracker was able to perform closer to the theoretical ideal of simply selecting what was looked at, a control form that can be very satisfying. Notably, though, even with a good calibration, the gaze tasks could frustrate the participant with undesirable selections while they looked around the board.

The perspective of the camera meant that some squares were larger than others. One participant noted: "Squares have different sizes in the video so some are difficult to choose by gaze." Another participant noted difficulty in selecting the lower squares with the eye gaze tracker. "The tracker worked well on some spaces but I had problems with the lower squares (maybe because of the angle?)." This indicates the perspective and accuracy of calibration through different parts of the screen may have had some influence on these selection issues. Regardless, it seems likely a top-down perspective would have achieved higher accuracy. Such a perspective is difficult to achieve, however, it seems reasonable to consider the angled view one of the limitations of dealing with a real device, an intended component of the experiment design. The natural tendency of the eye to track objects also created issues when in combination with some of the interface elements. If the gaze target was offset, participants found it difficult not to look at it, and then follow it as it shifted repeatedly due to the move of their gaze. Keeping the eyes steady also gave participants trouble. "Eyes have a natural tendency to move around, so it takes a lot of effort to keep them still and steady." Many also found keeping the head still problematic as well. Some participants mentioned that these factors were their reason for rating physical demand above minimum in their NASA TLX questionnaire.

It is interesting to note that, despite the number of issues that plagued the gaze-only system, it was ranked as best/easiest to use by a non-trivial number of participants. Two possible explanations for the high scores given by these participants are: "The challenge of controlling the eye gaze tracker and dealing with its inaccuracies makes it fun to use." and "The rapid responsiveness and feeling of it acting as I think found with the eye gaze tracker alone was so appealing it made up for the inaccuracies and other issues." The latter comment could imply that speed and responsiveness is enormously important to some participants. One of those who preferred the gaze commented "I preferred the eye gaze tracker to the controller (when calibrated well) because it used much less effort and was faster," indicating the importance of intuitive user interfaces that lower cognitive load. Such feedback warrants optimism that, when the gaze tracking works accurately and reliably, it is a very appealing interaction method.

It is also notable that several participants mentioned that they felt they were learning to use the gaze system significantly better throughout the study. Such comments include: "The gaze tracking became easier over time." and "The eye gaze tracker took some learning. The effort required decreased from task to task, even though its complexity increased." Viewing participant behaviour showed them adapting in a number of ways. Some would begin to adopt the suggestions given by the script to prevent eye gaze tracker selections, closing or covering their eyes. Others realised they could cover the eye gaze tracker instead of their own eyes to achieve the same effect. Often participants would begin to turn their head when the gaze was inaccurate, despite having been advised against this. While measures, such as the practice task and reordering of device use were taken to minimise the effect of learning on the experiment, previous experience most likely still had an impact. To overcome this limitation, longitudinal studies over a longer time frame would need to be performed for participants to familiarise themselves with the gaze tracker. In a past study, König et al. [31] compared multimodal feedback with conventional visual systems, finding the multimodal systems both performed better and saw greater improvement over the eight-day trial. While intuitiveness is highly desirable in interaction devices, in most real world scenarios, users spend enough time to become very familiar with their interfaces, thus results from experienced participants would be highly relevant.

6. Conclusions

Twenty participants completed an experiment that compared three interaction modalities for robot arm control: controller, gaze, and multimodal. The system involved control of a robot arm through a video display in order to ground the experiment in reality. The experimental setup allowed examination of how well a multimodal system performed relative to the individual modalities that compose it in a scenario limited by the practicalities of controlling a robot arm in the real world. Comparing the multimodal setup to gaze measured how well it performed at mitigating gaze interaction weaknesses. Comparing the multimodal setup to the controller measured how well it performed relative to a commonly used device. Finally, solving a number of chess tasks at various complexities allowed measurements of how the different interaction setups scaled with task complexity.

This research has shown that a multimodal setup can be effective at mitigating the weaknesses of gaze based interaction. The multimodal setup was faster for two of the three tasks, although not significantly. It was significantly better in terms of workload for all three tasks. The use of a multimodal setup is therefore an additional method through which the considerable potential advantages of eye gaze trackers can be accessed. The general effectiveness of such a multimodal setup was compared against a typical controller setup and found to be competitive. This indicates that there is some potential in gaze and controller multimodal setups as a multimodal system.

The results of this research also highlight the potential of eye gaze trackers as a method of interaction. While it might be expected that inaccuracy and the "Midas touch" issue would universally put users off this system, this study found a number of participants still preferred the gaze only setup. This setup was not designed to mitigate either of these issues, so the participants must have felt the benefits of the gaze system outweighed even these considerable problems. Examining this viewpoint in terms of both how many share it and the exact reasons they take this view offers an interesting avenue for further research. A study including experienced eye gaze control participants could be performed to see how experience impacts on overall performance. If they were able to achieve significantly better performance, it is expected that in the future eye gaze and multimodal control will become a very useful interaction method. More importantly, the method presented in out experiment used only one of the control modalities at a time within a task. However, a control methodology that could combine the inputs from the two modalities probabilistically can result in the most effective interaction interface. Future investigations involving this new method could result in more promising options for control of robot arms.

It should be noted that our study has limitations. For example, more participants would help to achieve more reliable statistical results. Further, all participants were from the same student population and they had an unbalanced experience with the test modalities, which undoubtedly had given advantages to one modality over another. Future studies should recruit a larger sample of participants from the target user population. In addition, sufficient practice should be given so that participants have the same level of familiarity with all modality conditions.

Author Contributions: Conceptualization, W.C., U.E. and C.L.; Data curation, E.R.J.; Formal analysis, E.R.J.; Investigation, E.R.J., W.C., W.H., U.E. and C.L.; Project administration, W.C.; Supervision, W.C., U.E. and C.L.; Writing-original draft, E.R.J., W.C. and W.H.; Writing-review & editing, W.H.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Biswas, P. *Exploring the Use of Eye Gaze Controlled Interfaces in Automotive Environments*; Springer: Basel, Switzerland, 2016; Volume 14, pp. 77–89.
2. Asan, O.; Yang, Y. Using Eye Trackers for Usability Evaluation of Health Information Technology: A Systematic Literature. *JMIR Hum. Factors* **2015**, *2*, e5. [CrossRef] [PubMed]
3. Huang, W. Establishing aesthetics based on human graph reading behavior: Two eye tracking studies. *Pers. Ubiquitous Comput.* **2013**, *17*, 93–105. [CrossRef]
4. Duchowski, A.T. Gaze-based interaction: A 30 year retrospective. *Comput. Graph.* **2018**, *73*, 59–69. [CrossRef]
5. He, H.; She, Y.; Xiahou, J.; Yao, J.; Li, J.; Hong, Q.; Ji, Y. Real-Time Eye-Gaze Based Interaction for Human Intention Prediction and Emotion Analysis. In Proceedings of the Computer Graphics International 2018, Bintan Island, Indonesia, 11–14 June 2018; ACM: New York, NY, USA, 2018; pp. 185–194. [CrossRef]
6. Bieg, H.J.; Chuang, L.L.; Fleming, R.W.; Reiterer, H.; Bülthoff, H.H. Eye and pointer coordination in search and selection tasks. In Proceedings of the 2010 Symposium on Eye Tracking Research & Applications, Austin, TX, USA, 22–24 March 2010; pp. 89–92.
7. Koch, C.; Ullman, S. Shifts in selective visual attention: Towards the underlying neural circuitry. *Hum. Neurobiol.* **1985**, *4*, 219–227. [PubMed]
8. Betz, T.; Kietzmann, T.C.; Wilming, N.; König, P. Investigating task-dependent top-down effects on overt visual attention. *J. Vis.* **2010**, *10*, 15. [CrossRef] [PubMed]
9. Gegenfurtner, K.R. The Interaction Between Vision and Eye Movements. *Perception* **2016**, *45*, 1333–1357. [CrossRef] [PubMed]
10. Borji, A.; Itti, L. Defending Yarbus: Eye movements reveal observers' task. *J. Vis.* **2014**, *14*, 29. [CrossRef] [PubMed]
11. Kanan, C.; Nicholas, R.A.; Bseiso, D.N.; Hsiao, J.H.; Cottrell, G.W. Predicting an observer's task using multi-fixation pattern analysis. In *Proceedings of the Symposium on Eye Tracking Research and Applications*; ACM: New York, NY, USA, 2014; pp. 287–290.
12. Stuart, S.; Alcock, L.; Godfrey, A.; Lord, S.; Rochester, L.; Galna, B. Accuracy and re-test reliability of mobile eye-tracking in Parkinson's disease and older adults. *Med. Eng. Phys.* **2016**, *38*, 308–315. [CrossRef] [PubMed]
13. Ziv, G. Gaze Behavior and Visual Attention: A Review of Eye Tracking Studies in Aviation. *Int. J. Aviat. Psychol.* **2016**, *26*, 75–104. [CrossRef]
14. Fernandez, D.R.; Niu, J.; Lochner, M. Fast Human-Computer Interaction by Combining Gaze Pointing and Face Gestures. *TACCESS* **2017**, *10*, 10:1–10:18.

15. Stellmach, S.; Dachselt, R. Still looking: Investigating seamless gaze-supported selection, positioning, and manipulation of distant targets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2013; pp. 285–294.
16. Instance, H.; Bates, R.; Hyrskykari, A.; Vickers, S. Snap clutch, a moded approach to solving the Midas touch problem. In Proceedings of the 2008 Symposim on Eye Tracking Research & Applications, Savannah, GA, USA, 26–28 March 2008; ACM: New York, NY, USA, 2008; pp. 221–228.
17. Velichkovsky, B.B.; Rumyantsev, M.A.; Morozov, M.A. New Solution to the Midas Touch Problem: Identification of Visual Commands Via Extraction of Focal Fixations. *Procedia Comput. Sci.* **2014**, *39*, 75–82. [CrossRef]
18. Morimoto, C.H.; Mimica, M.R. Eye gaze tracking techniques for interactive applications. *Comput. Vis. Image Underst.* **2004**, *98*, 4–24. [CrossRef]
19. Tuisku, O.; Surakka, V.; Vanhala, T.; Rantanen, V.; Lekkala, J. Wireless Face Interface: Using voluntary gaze direction and facial muscle activations for human-computer intraction. *Interact. Comput.* **2012**, *24*, 1–9. [CrossRef]
20. Bolt, R.A. "Put-that-there": Voice and gesture at the graphics interface. *Comput. Graph.* **1980**, *14*, 262–270. [CrossRef]
21. Turk, M. Multimodal interaction: A review. *Pattern Recognit. Lett.* **2014**, *36*, 189–195. [CrossRef]
22. Paas, F.; Renkl, A.; Sweller, J. Cognitive load theory and instructional design: Recent developments. *Educ. Psychol.* **2003**, *38*, 1–4. [CrossRef]
23. Boekaerts, M. Cognitive load and self-regulation: Attempts to build a bridge. *Learn. Instr.* **2017**, *51*, 90–97. [CrossRef]
24. Huang, W.; Luo, J.; Bednarz, T.; Duh, H. Making Graph Visualization a User-Centered Process. *J. Vis. Lang. Comput.* **2018**, *48*, 1–8. [CrossRef]
25. Huang, W.; Eades, P.; Hong, S.H. Measuring effectiveness of graph visualizations: A cognitive load perspective. *Inf. Vis.* **2009**, *8*, 139–152. [CrossRef]
26. Oviatt, S.; Coulston, R.; Lunsford, R. When do we interact multimodally? Cognitive load and multimodal communication patterns. In Proceedings of the 6th International Conference on Multimodal Interfaces, State College, PA, USA, 13–15 October 2004.
27. Hinkelmann, K. *Design and Analysis of Experiments*; Wiley: Hoboken, NJ, USA, 2008.
28. Hart, S.G. NASA-task load index (NASA-TLX); 20 years later. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2006**, *50*, 904–908. [CrossRef]
29. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Adv. Psychol.* **1988**, *52*, 139–183.
30. Hoste, L.; Dumas, B.; Signer, B. Speeg: A multimodal speech and gesture-based text input solution. In Proceedings of the International Working Conference on Advanced Visual Interfaces, Capri Island, Italy, 22–25 May 2012; pp. 156–163.
31. König, W.A.; Radle, R.; Reiterer, H. Interactive design of multimodal user interfaces. *J. Multimodal User Interfaces* **2010**, *3*, 197–213. [CrossRef]