Keyboard before Head Tracking Depresses User Success in Remote Camera Control

Dingyun Zhu[1,2], Tom Gedeon[2], and Ken Taylor[1]

[1] CSIRO ICT Centre, Acton, Canberra, ACT 0200, Australia
[2] School of Computer Science, College of Engineering and Computer Science, The Australian National University, Canberra, ACT 0200, Australia
dingyun.zhu@csiro.au, tom.gedeon@anu.edu.au, ken.taylor@csiro.au

Abstract. In remote mining, operators of complex machinery have more tasks or devices to control than they have hands. For example, operating a rock breaker requires two handed joystick control to position and fire the jackhammer, leaving the camera control to either automatic control or require the operator to switch between controls. We modelled such a teleoperated setting by performing experiments using a simple physical game analogue, being a half size table soccer game with two handles. The complex camera angles of the mining application were modelled by obscuring the direct view of the play area and the use of a Pan-Tilt-Zoom (PTZ) camera. The camera control was via either a keyboard or via head tracking using two different sets of head gestures called "head motion" and "head flicking" for turning camera motion on/off. Our results show that the head motion control was able to provide a comparable performance to using a keyboard, while head flicking was significantly worse. In addition, the sequence of use of the three control methods is highly significant. It appears that use of the keyboard first depresses successful use of the head tracking methods, with significantly better results when one of the head tracking methods was used first. Analysis of the qualitative survey data collected supports that the worst (by performance) method was disliked by participants. Surprisingly, use of that worst method as the first control method significantly enhanced performance using the other two control methods.

Keywords: Head Tracking, Remote Camera Control, Human Computer Interaction, Teleoperation, Usability Evaluation.

1 Introduction

Teleoperation has been regarded as an essential application strategy and widely applied in modern industry because of a variety of advantages. A device or machine is remotely operated by a person from a distance, which is able to effectively move human workers away from hazardous or difficult working environments, while potentially improving productivity and reducing costs. Regardless of whether the machine is directly manipulated by an operator, or granted full autonomy to execute its specific mission, at some level, human observation, supervision, and judgment remain critical elements of the entire teleoperation activity [7]. The direct perceptual link to the

T. Gross et al. (Eds.): INTERACT 2009, Part II, LNCS 5727, pp. 319–331, 2009.

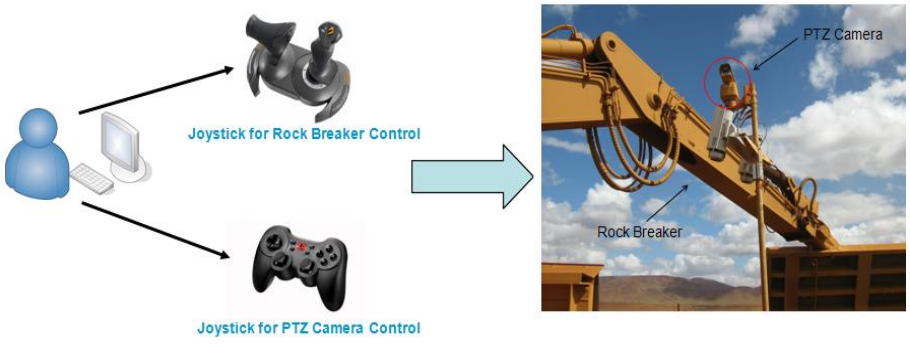

Fig. 1. Multi-task Situation for Rock Breaking in Remote Mining: for each device control, the operator has two joysticks, one hand does x, y and the other does z

remote environment often comes through a video feed supplied from one or more cameras as the foundation of situational awareness for the operator.

In practice, however, operators usually have to control multiple complex devices simultaneously, often more than they have hands, such as controlling a mechanical robot and a video camera at the same time. Figure 1 shows a typical teleoperation scenario for the rock breaking instance in mining. It is obvious that the operator is not capable of manipulating the two-joystick based 3D interfaces for the control of a rock breaker and to control a PTZ camera simultaneously as he can direct physical attention to only one task at a time. Under such circumstances, the operator needs to frequently switch hands between two different joysticks in order to accomplish the rock breaking task. This degrades the productivity of the entire process, increasing extra workload as well as the number of avoidable operational errors.

In this research, we focus on importing computer vision technology to undertake head tracking in interface design for teleoperation activities. The common remote control situation described above is modelled by using a physical game analogue: playing a table soccer game with two handles. This has the advantage of being more compelling for our student experimental subjects than a more abstract task. We use student experimental subjects as we have limited access to the operators. We then propose a novel design applying natural human head gestures for controlling a Pan-Tilt-Zoom camera as an effective approach to solve the camera control problem.

2 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 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.

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 2) for visual feedback and interaction.

There are a number of sensor based head trackers commercially available. TrackIR (Figure 3) is a typical head tracking device currently quite popular amongst gamers, especially in the simulation community [10]. 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. Head orientation can then be used as 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 [14].

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 device, and less cost for the hardware as usually only a normal webcam is needed.

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 [2]. In addition, color information is another powerful cue for locating the face [6] and other methods such as the use of depth information [8], classification of the brightness pattern inside an image window [12], etc. Figure 4 illustrates a commercialized real-time face tracking technology: FaceAPI, 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 [13].

Fig. 2. A Goggle Display for Head Tracking

Fig. 3. TrackIR System

Fig. 4. FaceAPI: Real-time Head Tracking with a Single Webcam

3 Head Tracking Applications in Human Computer Interaction

Head tracking is a key component in applications such as human computer interaction, person monitoring, driver monitoring, video conferencing, and object-based compression. Recently, one of the most popular ways of applying head tracking 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 [17], head tracking has been integrated into a first-person-shooter (FPS) game "Bullet Time" to control the user's view point. Another similar application with exaggerated head motions for game viewpoint control can also be found in [15].

In addition, there have also been attempts to develop head tracking based "hands-free pointing" interface for controlling the mouse cursor [16], by which a user can point his nose where he wishes to place the cursor on a monitor screen. "hMouse" is another head tracking driven camera mouse system [4], which provides alternative solutions for convenient device control with potential applications for people with disabilities and the elderly.

Other relevant applications are head tracking based user interfaces for navigation in virtual environments, remote control of devices [1] and head gestures (e.g. "Nodding" and "Shaking") based perceptual interface [3] [9].

4 Our Design of Head Gesture Based Remote Camera Control

The basic function of a PTZ camera is to Pan, Tilt and Zoom. With various functional combinations, the operator can obtain flexible control of its movement. With the integration of head tracking techniques with PTZ camera functions, we propose two sets of simple head gestures as interactive methods for remote control.

The first method (called "motion") operates according to natural human head motion. As shown in Figure 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, 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 6 shows similar interaction for the head tilting. When the user tilts their head up or

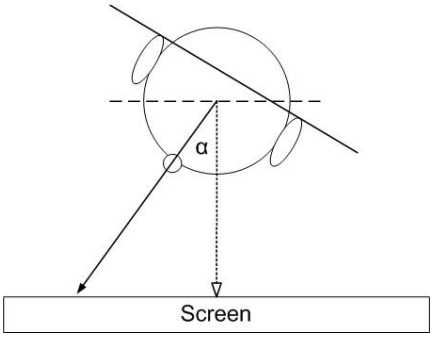

Fig. 5. Head Rotation

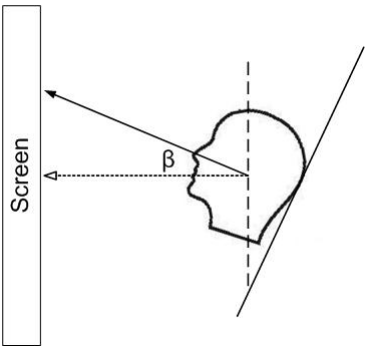

Fig. 6. Head Tilting

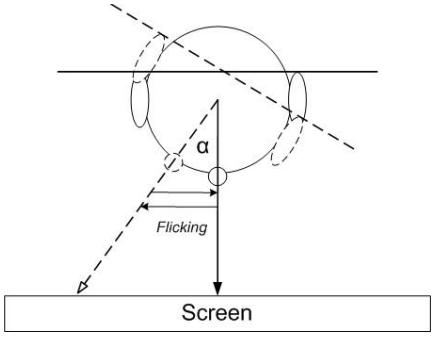

Fig. 7. Head Flicking for Panning

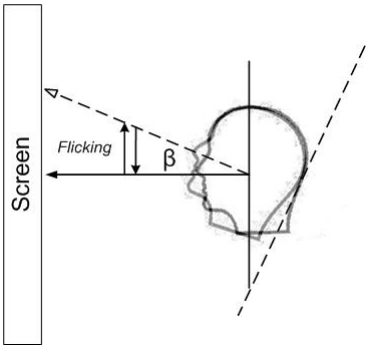

Fig. 8. Head Flicking for Tilting

down by a certain angle, the camera will correspondingly carry out the tilt function and not stop tilting until the head returns to the original position.

As for the zoom function, this specific control operates 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.

The other set of head gestures is based on human quick head movements, called "flicking". 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 7 and 8 are the relevant geometrical displays of head "flicking" for both pan right and tilt up actions respectively.

5 User Studies

A 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.

5. 1 Apparatus and Implementation

We integrated FaceAPI 3.03 [13] with a Logitech webcam [5] into our prototype system in VC++ that ran at 50Hz on a PC for real-time head tracking. The system used a Pelco ES30C PTZ camera [11] 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 9 and 10 show the experimental setup from front and back respectively.

5. 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 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.

5. 3 Experimental Design and Procedure

The experiment was conducted using a 3 × 2 × 3 × 3 within-subject full factorial 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* (S1: Motion → Flicking → Keyboard, S2: Keyboard → Motion → Flicking or S3: Flicking → Keyboard → Motion).

Please note in Figure 10, 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.

Particularly for the sequences, in order to avoid too many repeated trials which would affect the final results, a random selection of allocating participants into different experimental sequences was also carried out, which ended up with 5 subjects for the first sequence, 3 for the second and the remaining 2 for the last order.

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.

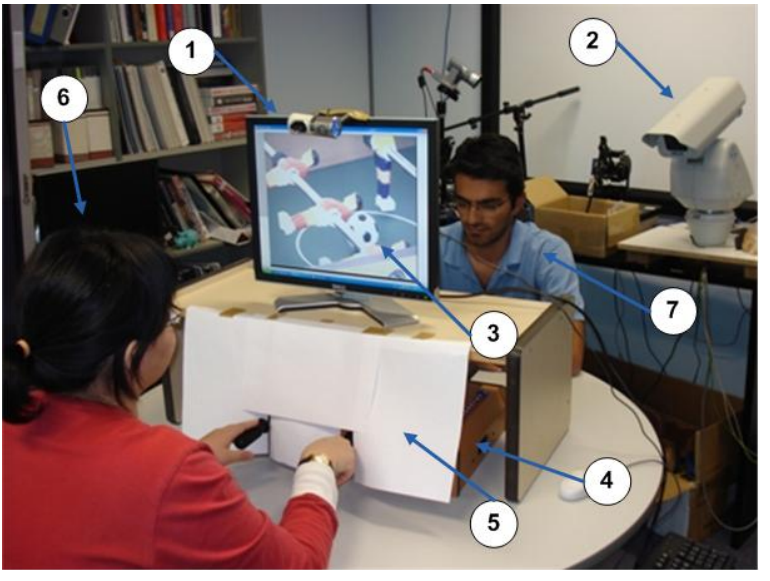

Fig. 9. Front View of the Test Setup, Webcam for Head Tracking (1), PTZ Camera (2), Video Stream from the PTZ Camera (3), Table Soccer (4), Covers for Obscuring Participant's Direct Vision (5), Experiment Participant (6), Experiment Assistant (7)

Fig. 10. Back View of the Test Setup

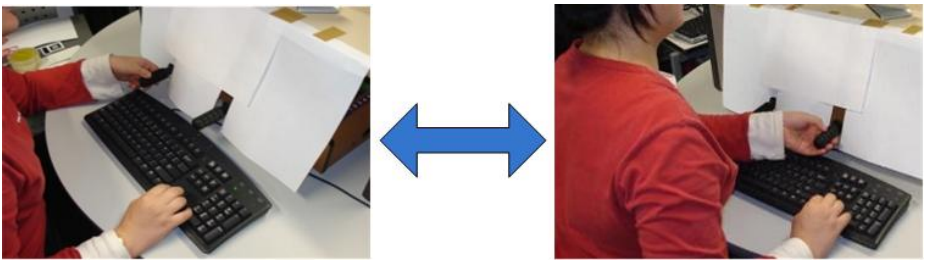

Fig. 11. One Hand Switching Between Handles, the Other Performing Camera Control by Keyboard

Participants were first given a short introduction (around 5 minutes) about the system, instructions on how to remotely control the PTZ camera by the two 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 time to get used to.

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 11). In the future, we also intend to test if restricting the user in hand switching makes a difference.

6 Experimental 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 12 shows the mean kicks for each 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 significant impact, F (2, 22) = 8.6814, $p < 0.01$. Participants who

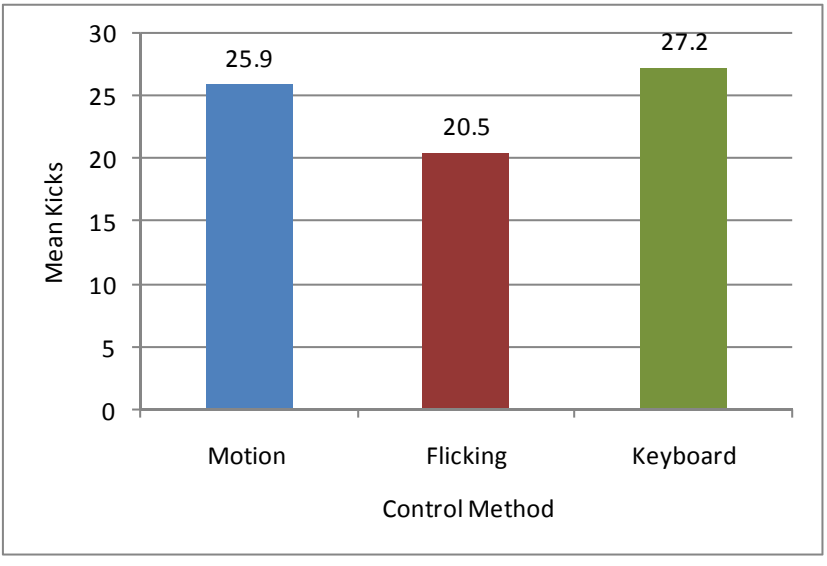

Fig. 12. Mean Kicks for Each Control Method

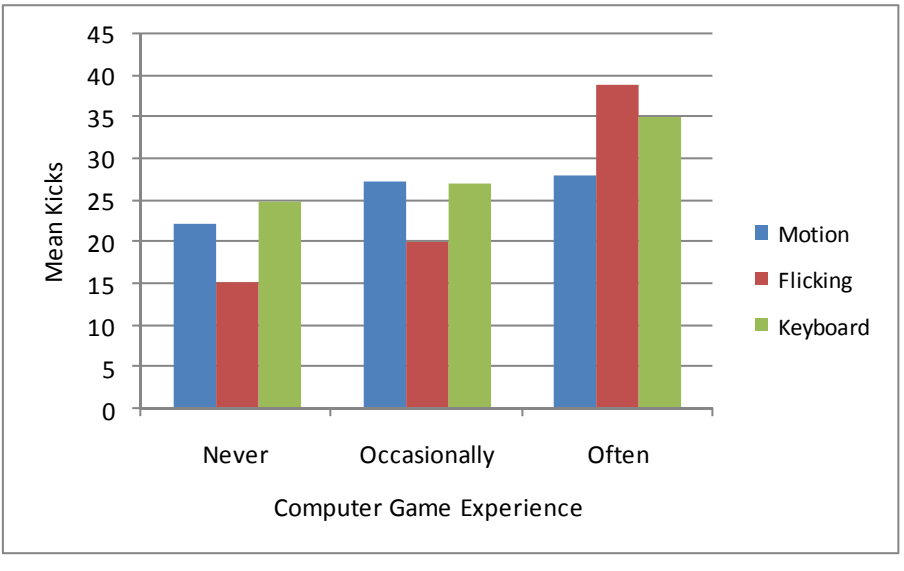

Fig. 13. Performance Comparison Based on Computer Game Experience

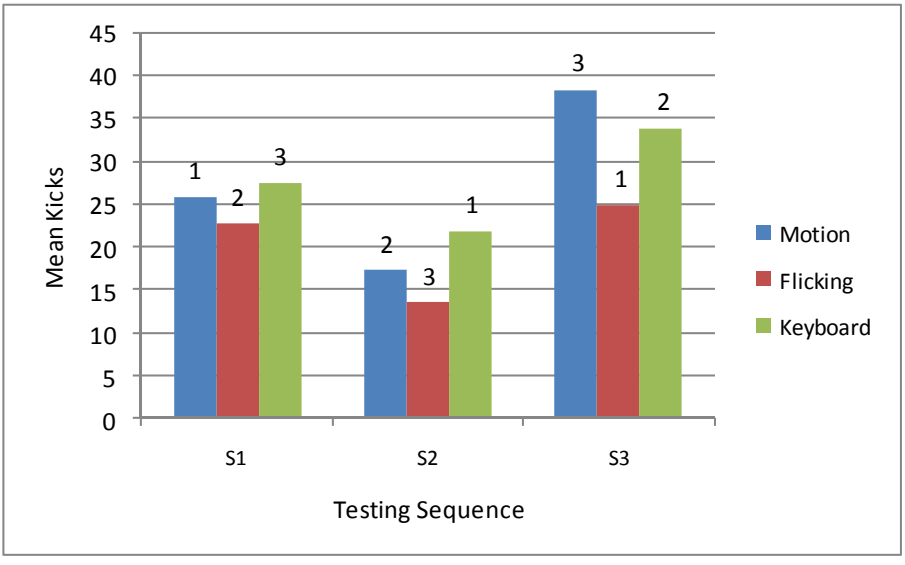

Fig. 14. Performance Comparison Based on Testing Sequence (S1: Motion → Flicking → Keyboard, S2: Keyboard → Motion → Flicking, S3: Flicking → Keyboard → Motion)

often played computer games using keyboards outperformed subjects with only occasional experience or no experience through all three different control conditions (see Figure 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 14).

Table 1 illustrates the average scores participants rated for these three different 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 1. User Preference Results from the Subjective Measure

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

7 Discussion

Our objective results indicate that for this specific experimental setting, keyboards still performed the best by 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.

Unsurprisingly, the "head flicking" strategy did not perform as well as the other head gesture 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 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 ($M_{\text{Keyboard}} = 22 > M_{\text{Motion}} = 17.33 > M_{\text{Flicking}} = 13.67$). 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 ($M_{\text{Keyboard}} = 25$, $M_{\text{Flicking}} = 38.5 > M_{\text{Motion}} = 34$). 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. $M_{\text{S1-Motion}} = 26 > M_{\text{S2-Motion}} = 17.33$, and $M_{\text{S1-Flicking}} = 22.8 > M_{\text{S2-Flicking}} = 13.67$; while $M_{\text{S3-Motion}} = 38.5 > M_{\text{S2-Motion}} = 17.33$, and $M_{\text{S3-Flicking}} = 25 > M_{\text{S2-Flicking}} = 13.67$).

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.

8 Conclusion

In this paper, we considered the common problem of requiring an operator to control multiple devices simultaneously in current teleoperation, especially in remote mining. We presented our approaches of using two different sets of human head gestures to control a PTZ camera as potential solutions for this real-world situation.

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 head motion control, head flicking control and keyboard control.

From the results, we demonstrate that the head motion based control is able to provide a comparable performance to using a keyboard even without the requirement of pre-training time, and the subjective measure of user's preference also indicates that the head motion is a comparable and effective method for this remote camera control case. Furthermore, we find that the sequence of conducting the three methods is the most significant factor. The use of keyboard control first depresses 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 warrants further investigation.

We believe our results map back to the mining teleoperation setting as follows. The results as to the two forms of head motion based control are likely to be directly applicable, so we expect flicking to be worst. The keyboard was familiar to our subjects, which most likely maps to joystick control in the mining setting, as the operators use joysticks a lot to control the rock breaker.

Our sequence results mapped to the mining setting would mean that operators presented with a teleoperation interface and initially presented with a joystick based control of the camera would have diminished performance with head motion based control. We could explain this by arguing that in the task encumbered mining setting, as long as the control of the camera is "good enough" minimal extra effort is expended on any later control strategy, as in real use the user knows they can resile to the "good enough" strategy. On the other hand, when initially presented with a novel interface, some effort is expended in learning to use it, which then has benefits for subsequent performance.

This explanation has two testable consequences when mapped back to our experimental setting. Firstly, that the same experiment with a less encumbered task would reduce or eliminate the effect of the keyboard in depressing head tracking based control. Secondly, that our physical model of the mining setting was sufficiently engaging to elicit such focus on the task, hence an equally encumbered but less engaging model would again reduce the effect of initial use of the keyboard.

Acknowledgements

The authors would like to express their appreciations to all the students and staff that participated in the experiment, and also thank Matt Adcock, Chris Gunn and Amir

Hadad for their great help on the implementation as well as the special assistance in the experiment.

Reference

1. Avizzano, C.A., Sorace, P., Checcacci, D., Bergamasco, M.: A Navigation Interface Based on Head Tracking by Accelerometers. In: 13th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN 2004), Kurashiki, Okayama, Japan, pp. 625–630 (2004)
2. Birchfield, S.: An Elliptical Head Tracker. In: 31st Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, vol. 2, pp. 1710–1714 (1997)
3. Davis, J.W., Vaks, S.: A Perceptual User Interface for Recognizing Head Gesture Acknowledgements. In: 2001 Workshop on Perceptive user interfaces (PUI), Orlando, Florida, USA, pp. 1–7 (2001)
4. Fu, Y., Huang, T.S.: hMouse: Head Tracking Driven Virtual Computer Mouse. In: 8th IEEE Workshop on Applications of Computer Vision (WACV 2007), p. 30 (2007)
5. Logitech, Inc. (2009), http://www.logitech.com/
6. Hsu, R.-L., Abdel-Mottaleb, M., Jain, A.K.: Face Detection in Color Images. IEEE Trans. Pattern Anal and Mach. Intel. 24(5), 696–706 (2002)
7. Hughes, S., Lewis, M.: Robotic Camera Control for Remote Exploration. In: ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2004), Vienna, Austria, pp. 511–517 (2004)
8. Malassiotis, S., Strintzis, M.G.: Real-time Head Tracking and 3D Pose Estimation from Range Data. In: 2003 International Conference on Image Processing (ICIP), Barcelona, Spain, vol. 2, pp. 895–862 (2003)
9. Morency, L.-P., Sidner, C., Lee, C., Darrel, T.: Contextual Recognition of Head Gestures. In: 7th International Conference on Multimodal Interfaces (ICMI 2005), Torento, Italy, pp. 18–24 (2005)
10. Natural Point, Inc.: TrackIR (2009), http://www.naturalpoint.com
11. Pelco, Inc. (2009), http://www.pelco.com/
12. Rowley, H.A., Baluja, S., Kanade, T.: Neural Network-based Face Detection. IEEE Trans. Pattern Anal and Mach. Intel. 20, 23–38 (1998)
13. Seeingmachines, Inc: FaceAPI (2009), http://www.seeingmachines.com/faceAPI.html
14. Smith, J.D., Nicholas Graham, T.C.: Use of Eye Movements for Video Game Control. In: 2006 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE 2006), California, USA, p. 20 (2006)
15. Teather, R.J., Stuerzlinger, W.: Exaggerated Head Motions for Game Viewpoint Control. In: ACM International Academic Games Conference on the Future of Game Design and Technology (FuturePlay 2008), Toronto, Ontario, Canada, pp. 240–243 (2008)
16. Toyama, K.: Look, Ma – No Hands! Hands-Free Cursor Control with Real-Time 3D Face Tracking. In: Workshop on Perceptual User Interfaces (PUI 1998), San Fransisco, USA, pp. 49–54 (1998)
17. Wang, S., Xiong, X., Xu, Y., Wang, C., Zhang, W., Dai, X., Zhang, D.: Face Tracking as an Augmented Input in Video Games: Enhancing Presence, Role-playing and Control. In: ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2006), Montréal, Québec, Canada, pp. 1097–1106 (2006)