A Pilot Study on Gaze-Based Control of a Virtual Camera Using 360°-Video Data

Jutta Hild[1](✉), Edmund Klaus[1], Jan-Hendrik Hammer[1], Manuel Martin[1], Michael Voit[1], Elisabeth Peinsipp-Byma[1], and Jürgen Beyerer[1,2]

[1] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Karlsruhe, Germany
jutta.hild@iosb.fraunhofer.de

[2] Vision and Fusion Laboratory, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Abstract. Over the last decades, gaze input appeared to provide an easy to use and less demanding human-computer interaction method for various applications. It appeared to be particularly beneficial in situations where manual input is either not possible or is challenging and exhausting like interaction with dynamic content in video analysis or computer gaming. In this contribution is investigated whether gaze input could be an appropriate input technique for camera control (panning and tilting) without any manual intervention. The main challenge of such an interaction method is to relieve the human operator from consciously interacting and to let them deploy their perceptive and cognitive resources completely to scene observation. As a first step, a pilot study was conducted operationalizing camera control by navigating in a virtual camera scene, comparing gaze control of the camera with manual mouse control. The experimental task required the 28 subjects (18 expert video analysts, 10 students and colleagues) to navigate in a 360° camera scene in order to keep track of certain target persons. Therefore, an experimental system was implemented providing virtual camera navigation in previously recorded 360fly camera imagery. The results showed that subjects rated gaze control significantly less loading than manual mouse control, using the NASA-TLX questionnaire. Moreover, the large majority preferred gaze control over manual mouse control.

Keywords: Virtual camera control · 360° video · Gaze input
Surveillance task

1 Introduction

Gaze-based control of user interfaces has been proposed and evaluated in plenty of contributions addressing various application domains. Researchers investigated gaze input for common desktop interaction tasks like object selection, eye typing, or password entry [1–3], for zooming maps or windows [4, 5], for foveated video streaming [6, 7], and PTZ camera remote control for surveillance [8] or teleoperation [9].

All implementations make use of gaze as a natural pointing »device«, as gaze is typically directed to the region of visual interest within the environment. Even though gaze has been evolved for perception, it has been shown that gaze can also be utilized

D. Harris (Ed.): EPCE 2018, LNAI 10906, pp. 419–428, 2018.
https://doi.org/10.1007/978-3-319-91122-9_34

as a method for information input. Particularly, gaze input is an alternative in situations where manual input is not possible, e.g., due to motor impairment [2] and in hands-busy and attention switching situations [9]. Moreover, gaze input proved to be a beneficial alternative for interaction in dynamic scenes, e.g., for moving target selection in full motion video [10], where manual input might be exhausting and challenging.

Recently, the eye tracking manufacturer Tobii started to make eye tracking and gaze interaction suitable for another application domain where interaction in dynamic scenes happens – the mass market of computer gaming. They provide the low-cost eye tracking device Tobii »4C« for 149$ (159€) [11]. Navigation in the scene of computer games using a first-person perspective is one of the proposed gaze input methods [12]. If the user directs their gaze, e.g., to the right corner of the current scene, the image section changes with the right corner subsequently becoming the next scene center. Thus, the visual focus of interest is brought to the scene center without any manual intervention. Such interaction models have also been proposed before by several authors investigating gaze input for computer gaming, e.g., [13, 14].

A similar kind of interaction is required when controlling a camera in a video surveillance task. Due to the rich visual input, this task can be very exhausting for the human operator, particularly, if the camera is mounted on a moving platform. Hence, any reduction of workload caused by less demanding human-computer interaction is welcome as it frees cognitive capacities for the actual surveillance task. A frequently occurring task is keeping track of a moving object, e.g., a person. If the object moves out of the currently displayed image section, the human operator must redirect the camera field of view. Gaze-based control of a camera appears compelling, keeping the camera focused on the object by just looking at this object. That way, the observer's visual attention could be focused on the (primary) surveillance task and the (secondary) interaction task is accomplished effortlessly at the same time.

In order to find out, whether such gaze interaction is appropriate and convenient, an experimental system was implemented simulating the control of a virtual camera as navigation in 360° video imagery. The system was evaluated in a user study with 28 participants, comparing gaze control versus manual control during the task of visually tracking a moving person.

2 Experimental System

The experimental system was implemented as a Java application which is able to play 360° video data recorded by the 360fly camera [15]. Figure 1 shows a video frame captured at an altitude of 30 m. For presentation to the user, the raw video data is rectified first, and in the next step, an image section of (width x height) 125° × 70° of the rectified 360° video data is provided on the user interface (Fig. 2). In related work, Boehm et al. [16] introduce a similar system displaying an image section of a 185° fisheye camera.

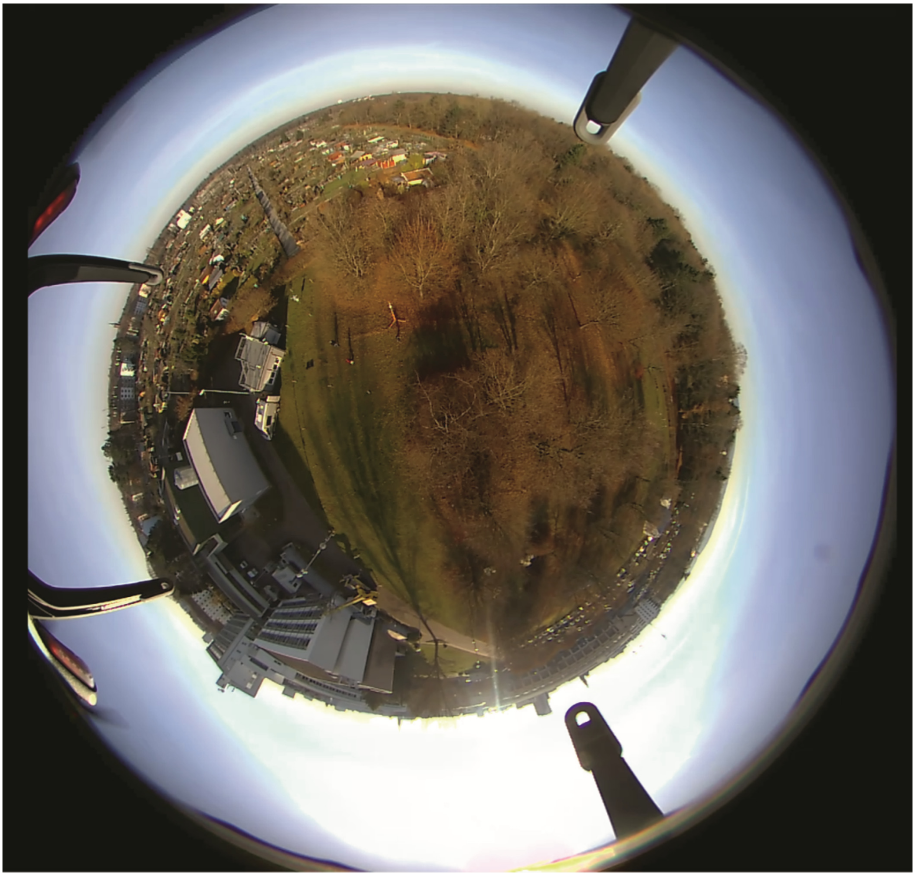

Fig. 1. Video data captured by 360fly at an altitude of 30 m.

Gaze interaction is performed using the Tobii 4C eye-tracker [10]. They provide gaze data in different modes [17]; in our system, the »lightly filtered« mode is used and passing additional low-pass filtering before being processed in the application. Figure 3 shows the underlying gaze interaction model for navigation in the scene. When the gaze position is located within the center region (white), the displayed image section remains the same and the human operator is enabled to calmly inspect that central region. When the gaze is located off the center region (blue), the image section is re-centered on this gaze position. The farther the eye gaze is directed away from the center towards the edges or corners, the faster the image section is centered on the new gaze position; similar models have been proposed before for remote camera control for surveillance [8] and teleoperation [9]. Calculation of the repositioning speed is based on the squared Euclidian distance between current gaze position and screen center. The maximum allowed speed for image section repositioning (achieved if looking at the edges) is 3° per frame (frame rate is 60 Hz).

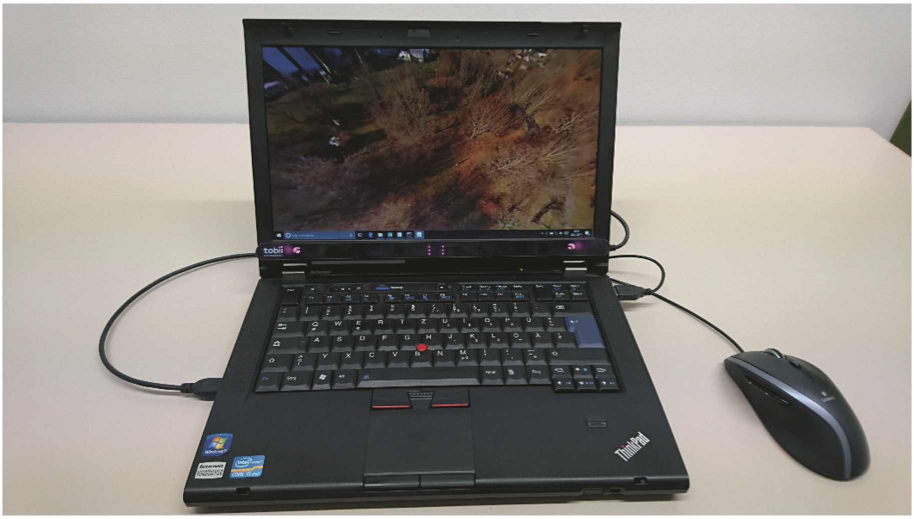

Fig. 2. Experimental system: image section displayed full-screen on a 14in laptop, equipped with a Tobii 4C eye-tracking device for gaze input, and a standard computer mouse providing the manual input alternative.

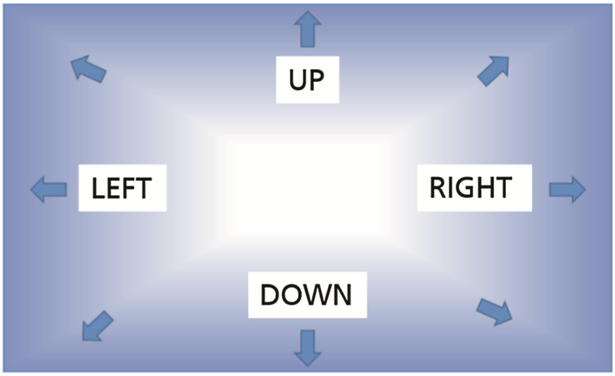

Fig. 3. The Gaze interaction model visualizing the activation dynamics on the screen: Gaze positions on the center region (white) have no effect. Gaze positions off the center region (blue) re-center the displayed image section to that gaze position; the closer to edges/corners (darker blue) a gaze position is located, the faster the image section is re-centered. (Color figure online)

The experimental system allows image section repositioning also with manual interaction using a computer mouse (Fig. 2). The user selects the image position of visual interest by pressing the left mouse button, then »drags« the image position with the left mouse button pressed to the wanted new position, for example the screen center.

3 Methodology

A pilot study was conducted to get first insight about the subjective workload of the gaze-based (virtual) camera control. 28 subjects (25 male, 3 female; 18 expert video analysts, 10 students and colleagues) performed the experimental task »Keep track of a person« using two different 3-min video sequences. Once, the test task instruction was to »Keep track of the person wearing the black jacket«, once »Keep track of the person wearing the red jacket« (Fig. 4). The video material was captured at an altitude of 30 m using a 360fly camera mounted on a 3DR solo drone [18]. The subjects were sitting at a distance of about 60 cm from the monitor (Fig. 5), the target persons' sizes therefore covered about 0.3° × 0.3° of visual angle on screen.

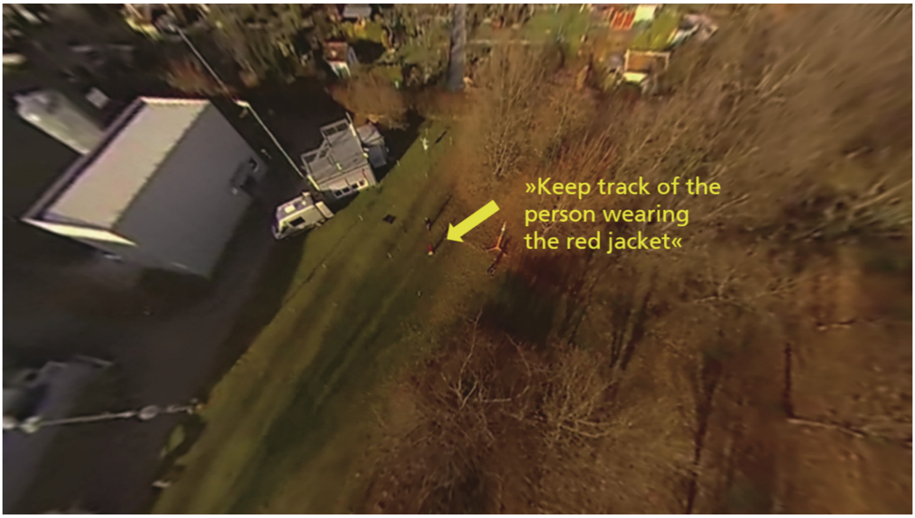

Fig. 4. Screenshot of a test ask with a target person. (Color figure online)

To ensure that subjects would have to reposition the scene in order to be able to keep track of the target person, the actors had been told to vary their motion trajectory and speed during video recording; thus, they temporarily moved straight on, or unpredictably, and sometimes shortly disappeared when walking under a tree. Furthermore, the drone and therefore also the camera trajectory carried out various motion patterns, like following an actor's trajectory, crossing an actor's trajectories, orbiting around the actors, or rotating at a stationary position. After performing the two test tasks, the subject answered the NASA-TLX [19, 20] questionnaire applied in the »Raw TLX« version, eliminating the weighting process.

For better interpretability of the NASA-TLX results for gaze input, the experimental design also incorporated performing the two test tasks using mouse input, and assessing it using the NASA-TLX. Half of the subjects performed the test tasks with gaze input

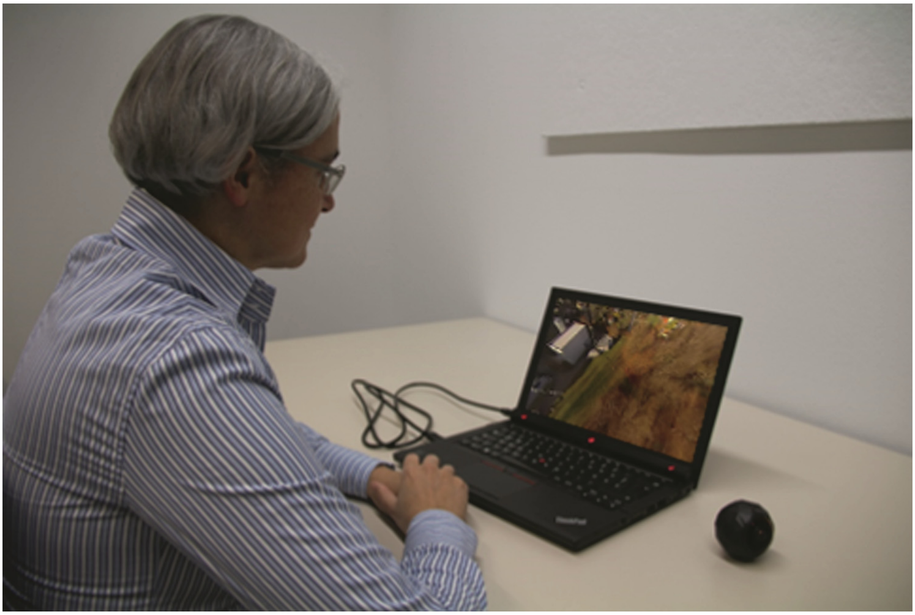

Fig. 5. Experimental setup.

first, the other half performed with mouse input first. The data recording of the 10 non-expert subjects was carried out in our lab, the data recording of the 18 expert video analysts was carried out at two locations of the German armed forces.

The procedure was as follows. Subjects were introduced into the experimental task but kept naïve in terms of the purpose of the investigation. Then, they performed the test tasks with the two interaction conditions one after another. In case of gaze input, subjects started performing the eye-tracker calibration provided by the Tobii-Software which requires fixating 7 calibration points; the calibration procedure was repeated until the offset between each fixated point and corresponding estimated gaze position was less than 1° of visual angle. Then, subjects got a different 3-min video sequence for training of the experimental task using that interaction technique. After that, subjects performed the two test tasks, immediately followed by rating their subjective workload using the NASA-TLX questionnaire. The mouse input condition was carried out performing the same three steps of training task, test tasks, and NASA-TLX rating. Finally, subjects were asked for their preferred interaction technique. The total duration of a session was about 30 min.

4 Results

The NASA-TLX results show that gaze input was rated with less workload both overall and in all single TLX categories. Results are provided using descriptive statistics as means with 1 standard deviation in Fig. 6 for all 28 subjects, in Fig. 7 for the expert video analysts only ($N = 18$). From those 18 experts, ten experts had much current practice in video surveillance and therefore were analyzed again, separately; results are shown in Fig. 8.

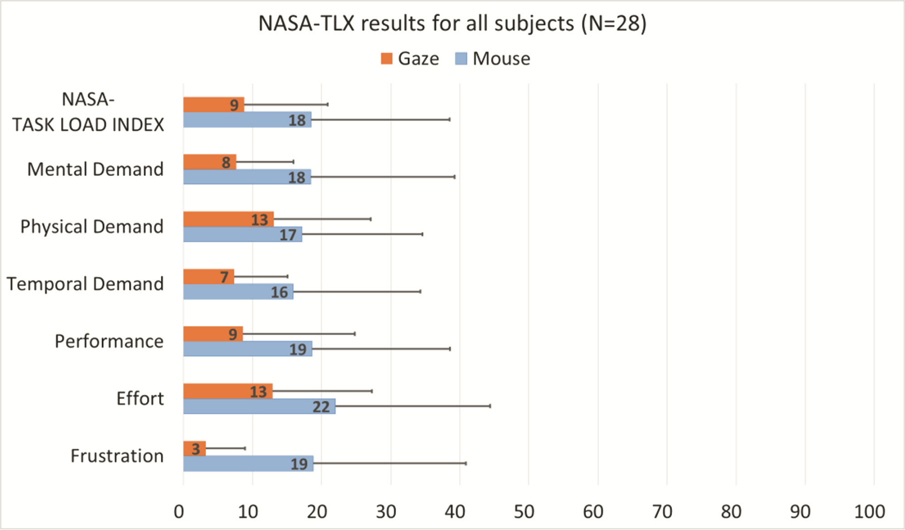

Fig. 6. Subjective workload with gaze input and mouse input, for all subjects.

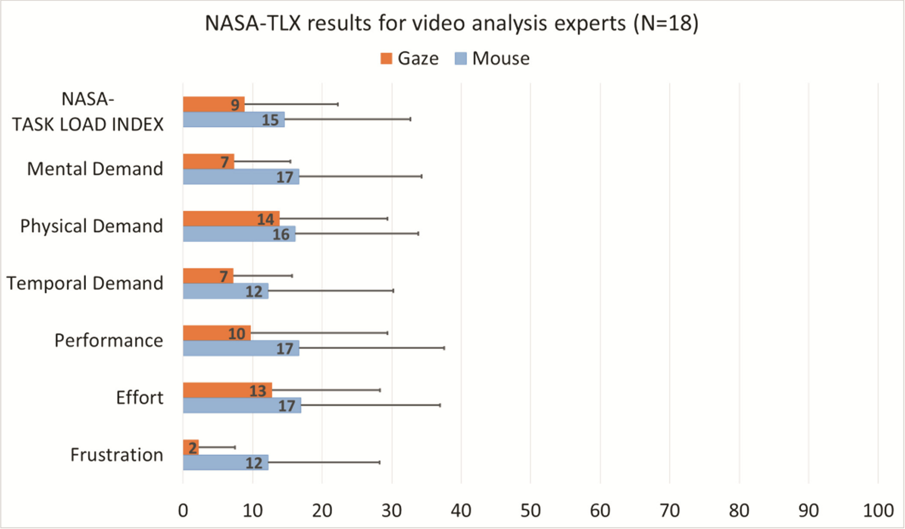

Fig. 7. Subjective workload with gaze input and mouse input, for subjects with expertise in video analysis.

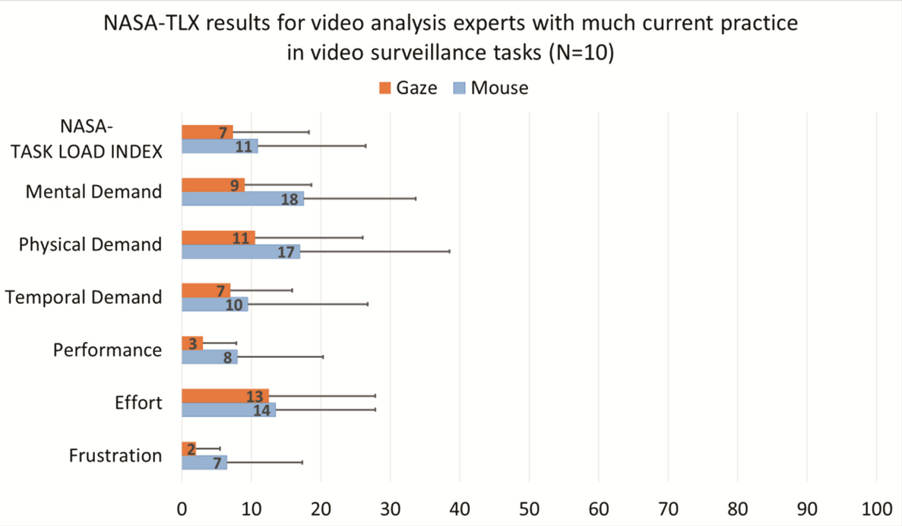

Fig. 8. Subjective workload with gaze input and mouse input, for subjects with expertise in video analysis and much current practice in video surveillance.

The NASA-TLX score is low for both interaction techniques, but it is significantly better for gaze input: A Wilcoxon signed-rank test for paired samples ($\alpha = 0.05$) revealed significant differences with $p < 0.001$ for $N = 28$, and $p < 0.05$ for $N = 18$; the result for the experts with much current practice in video surveillance ($N = 10$) is not significant ($p = 0.153$).

Analysis of the six subscales revealed further significant differences when analyzing all subjects ($N = 28$), for mental demand with $p < 0.05$, temporal demand $p < 0.01$, performance $p < 0.01$, effort $p < 0.05$, and frustration $p < 0.001$. Subscale analysis for expert video analysts ($N = 18$) and experts with much current practice in video surveillance ($N = 10$) still shows a significant difference between gaze and mouse for frustration ($p < 0.05$) despite the few data samples.

For mouse control, it can be observed that the subjective workload depends on video analysis expertise and current practice: The more expertise and practice, the lower the subjective workload (resulting in the NASA-TLX score difference between gaze and mouse being not significant any more, as reported above). However, for gaze control, subjective workload is very low for all subjects, independent of expertise. So, at least for control of a virtual camera, gaze input seems to be the more appropriate and convenient method to use.

Asked for their preference, 25 subjects preferred gaze input, 3 preferred mouse input ($N = 18$ experts: 15 preferred gaze input, 3 mouse input; $N = 10$ experts with much current practice in video surveillance: 10 preferred gaze input, 1 mouse input).

5 Conclusion

A pilot study was conducted in order to find out whether gaze input could be an appropriate input technique for camera control (panning and tilting) without any manual intervention. 28 subjects (18 expert video analysts from the German forces and 10 non-experts in video analysis) participated in the user study. Each performed the experimental task of tracking a target person using both gaze input and mouse input for navigation in a virtual camera, implemented based on 360° video imagery. The NASA-TLX showed that subjects rated both interaction conditions imposing rather little workload; however, gaze input was rated imposing significantly lower workload than mouse input. Hence, gaze input showed its potential to provide effortless interaction for this application, as it did for many other applications before.

Recently, the experimental system has been refactored and now besides navigation in recorded 360fly video data also allows live navigation in 360fly imagery. Future work will address gaze control for a real sensor, and user testing will show how workload would turn out to be in such condition with interaction latencies due to the necessary gimbal movements. Furthermore, future user studies will include more complex test tasks like observing more than one target object, as well as test tasks with a longer duration.

References

1. Jacob, R.J.: The use of eye movements in human-computer interaction techniques: what you look at is what you get. ACM Trans. Inf. Syst. (TOIS) **9**(2), 152–169 (1991)
2. Majaranta, P., Räihä, K.J.: Twenty years of eye typing: systems and design issues. In: Proceedings of the 2002 Symposium on Eye Tracking Research and Applications, pp. 15–22. ACM, New York (2002)
3. Kumar, M., Garfinkel, T., Boneh, D., Winograd, T.: Reducing shoulder-surfing by using gaze-based password entry. In: Proceedings of the 3rd Symposium on Usable Privacy and Security, pp. 13–19. ACM, New York (2007)
4. Fono, D., Vertegaal, R.: EyeWindows: evaluation of eye-controlled zooming windows for focus selection. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 151–160. ACM, New York (2005)
5. Stellmach, S., Dachselt, R.: Investigating gaze-supported multimodal pan and zoom. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 357–360. ACM, New York (2012)
6. Ryoo, J., Yun, K., Samaras, D., Das, S.R., Zelinsky, G.: Design and evaluation of a foveated video streaming service for commodity client devices. In: Proceedings of the 7th International Conference on Multimedia Systems. ACM, New York (2016)
7. Illahi, G., Siekkinen, M., Masala, E.: Foveated video streaming for cloud gaming. arXiv preprint arXiv:1706.04804 (2017)
8. Kotus, J., Kunka, B., Czyzewski, A., Szczuko, P., Dalka, P., Rybacki, R.: Gaze-tracking and acoustic vector sensors technologies for PTZ camera steering and acoustic event detection. In: Workshop on Database and Expert Systems Applications (DEXA), pp. 276–280. IEEE (2010)
9. Zhu, D., Gedeon, T., Taylor, K.: "Moving to the centre": a gaze-driven remote camera control for teleoperation. Interact. Comput. **23**(1), 85–95 (2010)

10. Hild, J., Kühnle, C., Beyerer, J.: Gaze-based moving target acquisition in real-time full motion video. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications, pp. 241–244. ACM, New York (2016)
11. Tobii Homepage. https://tobiigaming.com/eye-tracker-4c/. Accessed 05 Feb 2018
12. Tobii Homepage. https://tobiigaming.com/games/assassins-creed-origins/. Accessed 05 Feb 2018
13. Castellina, E., Corno, F.: Multimodal gaze interaction in 3D virtual environments. COGAIN **8**, 33–37 (2008)
14. Nacke, L.E., Stellmach, S., Sasse, D., Lindley, C.A.: Gameplay experience in a gaze interaction game. arXiv preprint arXiv:1004.0259 (2009)
15. 360fly Homepage. https://www.360fly.com/. Accessed 05 Feb 2018
16. Boehm, F., Schneemilch, S., Schulte, A.: The electronic camera gimbal. In: AIAA Infotech@Aerospace Conference (2013)
17. Tobii Homepage. https://tobii.github.io/CoreSDK/articles/streams.html. Accessed 05 Feb 2018
18. 3DR Homepage. https://3dr.com/solo-drone/. Accessed 05 Feb 2018
19. Hart, S.G.: NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, no. 9, pp. 904–908. Sage Publications, Los Angeles (2006)
20. NASA TLX Homepage. https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf. Accessed 05 Feb 2018