Automated interpretation of eye–hand coordination in mobile eyet racking recordings

Künstl Intell (2017) 31:331–337
DOI 10.1007/s13218-017-0503-y

TECHNICAL CONTRIBUTION

Automated interpretation of eye–hand coordination in mobile eye tracking recordings

Identifying demanding phases in human–machine interactions

Moritz Mussgnug[1] · Daniel Singer[1] · Quentin Lohmeyer[1] · Mirko Meboldt[1]

Received: 30 January 2017 / Accepted: 3 August 2017 / Published online: 17 August 2017

Abstract Mobile eye tracking is beneficial for the analysis of human–machine interactions of tangible products, as it tracks the eye movements reliably in natural environments, and it allows for insights into human behaviour and the associated cognitive processes. However, current methods require a manual screening of the video footage, which is time-consuming and subjective. This work aims to automatically detect cognitive demanding phases in mobile eye tracking recordings. The approach presented combines the user's perception (gaze) and action (hand) to isolate demanding interactions based upon a multi-modal feature level fusion. It was validated in a usability study of a 3D printer with 40 participants by comparing the usability problems found to a thorough manual analysis. The new approach detected 17 out of 19 problems, while the time for manual analyses was reduced by 63%. More than eye tracking alone, adding the information of the hand enriches the insights into human behaviour. The field of AI could significantly advance our approach by improving the hand-tracking through region proposal CNNs, by detecting the parts of a product and mapping the demanding interactions to these parts, or even by a fully automated end-to-end detection of demanding interactions via deep learning. This could set the basis for machines providing real-time assistance to the machine's users in cases where they are struggling.

Keywords Mobile eye tracking · Eye-hand coordination · Cognitive processes · Human–machine interaction · Event interpretation · Usability testing

1 Introduction

Eye–hand coordination plays an important role in many human–machine interactions with tangible products, e.g. using a coffee machine. In such real-world scenarios there is a striking and intimate link between eye movements and behavioural goals [11, 20].

Behavioural goals can be broken down into a series of actions directed towards various target objects [1], whereby the actions typically have durations of around 3 s [10]. Land and Tatler emphasised that “attention is intricately enmeshed in the task structure and actions that we engage in” [21]. Koenig et al. added that the eye movements “provide a window to the pacing [of actions] and the relevant variables of multistep behaviour ...” [9].

The eyes can be described as the slave to the motor system, as their function is to seek the information necessary to steer the actions [3]. Thus, decisions on the gaze location are linked to the needs of the current action. Hereby, the visual system acquires relevant information for fulfilling the subgoal just in time, instead of collecting those data beforehand [22]. Hence, analysing the eye–hand coordination can be directly assigned to the current behavioural sub-goal.

When analysing the dynamics of a sub-goal, the eyes and the motor system can be in one of two states: either both are focussed on the same target, i.e. the eyes are aiding the hands to perform the action, or the eyes move to other targets during the action, i.e. the motor system has sufficient information to operate without guidance.

At the start of a sub-goal, the gaze usually arrives at the corresponding target before the motor system starts to act, as it collects the relevant information to plan the motor system's movements [11]. Similarly, when the outcome of the current action can be predicted with confidence the gaze could already move to the target of the next phase [1].

✉ Moritz Mussgnug
mmussgnug@ethz.ch

[1] ETH Zurich, Leonhardstrasse 21, 8092 Zurich, Switzerland

This phenomenon is known as *look-ahead fixation* [6] and there is evidence that such fixations acquire information that is useful for the subsequent action [13]. As Tatler and Land stated,

> "At times when actions do not require strict monitoring, the attentional system can take advantage of these opportunities to look ahead and acquire information useful to the next part of the ongoing behaviour." [21]

Hence, depending on the demands of the current action, the gaze stays or already shifts to the next target [1]. Thus, only during *focussed interaction phases* the gaze and the hand both are rigidly directed to the target of the current action.

Mobile eye tracking (MET) describes wearable systems tracking the user's gaze [4]. Recent advancements in this technology make it possible to record human behaviour in almost any real-world settings in a non-invasive and unobtrusive manner [7, 21]. The scene camera of the eye tracker, which records the user's field of vision in the first-person perspective, also captures the user's actions performed with the hands, if the head is directed to that action. According to the eye-mind hypothesis, the location gazed upon is connected to what is simultaneously processed in the mind [8]. The gaze does not directly allow cognition to be measured, but can act as a window to cognitive processes [9].

Being able to measure the visual attention allows for new insights at the intersection of human behaviour and artificial intelligence, e.g. in a semantic interpretation of dynamic visuo-spatial imagery [19] or in the analysis of cognitive workload induced by artificial intelligence systems [2]. The work of [19] deals with spacio-temporal relations of objects detected in videos, which is relevant in the analysis of eye-hand coordination as well.

MET is a suitable tool to analyse the usability of tangible products [15], which are understood as physical objects operated with the hands. However, the analysis of MET videos is time-consuming and subjective because it is performed exclusively manually, which is a key obstacle to make this process efficient [5].

2 Research goal

The goal of this investigation is to automatically detect cognitive demanding handling interactions in long MET videos, with durations of 5–19 min per participant. Therefore we distinguish between two states: cognitive demanding and fluent. An interaction is seen as cognitive demanding if a user does not know how to approach a task or if the handling itself is difficult. To isolate demanding interactions, two behavioural aspects are combined in a multi-modal feature level fusion and their spacio-temporal relation is analysed.

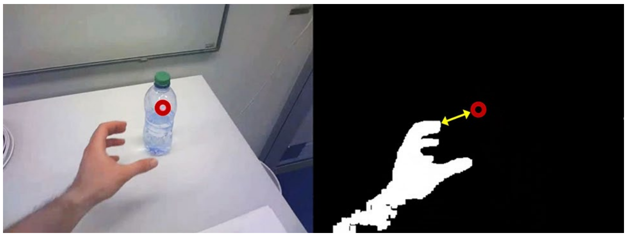

Fig. 1 *Left* Original frame from the scene video; *right* applied colour segmentation. The *red circle* shows the gaze location and the *yellow arrow* represents the hand-gaze distance. (Color figure online)

- The focus of the perception is considered by the location of the gaze, taken from the MET raw data.
- The focus of the action is considered by the position of the hand, extracted from the MET scene video.

We hypothesise that cognitive demanding handling interactions are represented by *long* periods of *constant* hand-gaze distance, as in these phases the hand and the gaze are involved in the same action.

3 Approach

This section describes how demanding human–machine interactions are isolated from MET data[1] in three parts: the detection of the hands, the identification of focussed interactions and the deduction of usability problems.

Detection of the hands To locate the position of the hand in the ego-centric scene video a pixel-wise classification through colour detection is applied [12]. Two colour spaces (HSV and RGB) are combined with a simple hue segmentation, as described by Song et al. [18]. Aiming to detect closed areas which most accurately represent the hands, two post-processing steps are performed. An erosion-and-dilation step eliminates most of the noise. Only areas larger than 15,000 px then are selected as hands. An example is presented in Fig. 1.

Identification of focussed interactions The hand-gaze distance (HGD) is acquired by measuring the minimum distance between the detected hand area and the gaze location (Fig. 1, yellow arrow). To isolate the focussed interactions on the basis of the HGD, four processing steps are performed (cf. Fig. 2).

In **Step 1**, the raw HGD signal is smoothed with a rolling median filter (window size 10). Since the focussed

[1] The investigation is based on the *SMI Eye Tracking Glasses 2w*, providing a scene camera resolution of 960 by 720 px (60° horizontal, 46° vertical) at 30 Hz. This refresh rate is interpolated to match the 60 Hz of the eye movement signal, which has a precision of 0.5° over all distances.

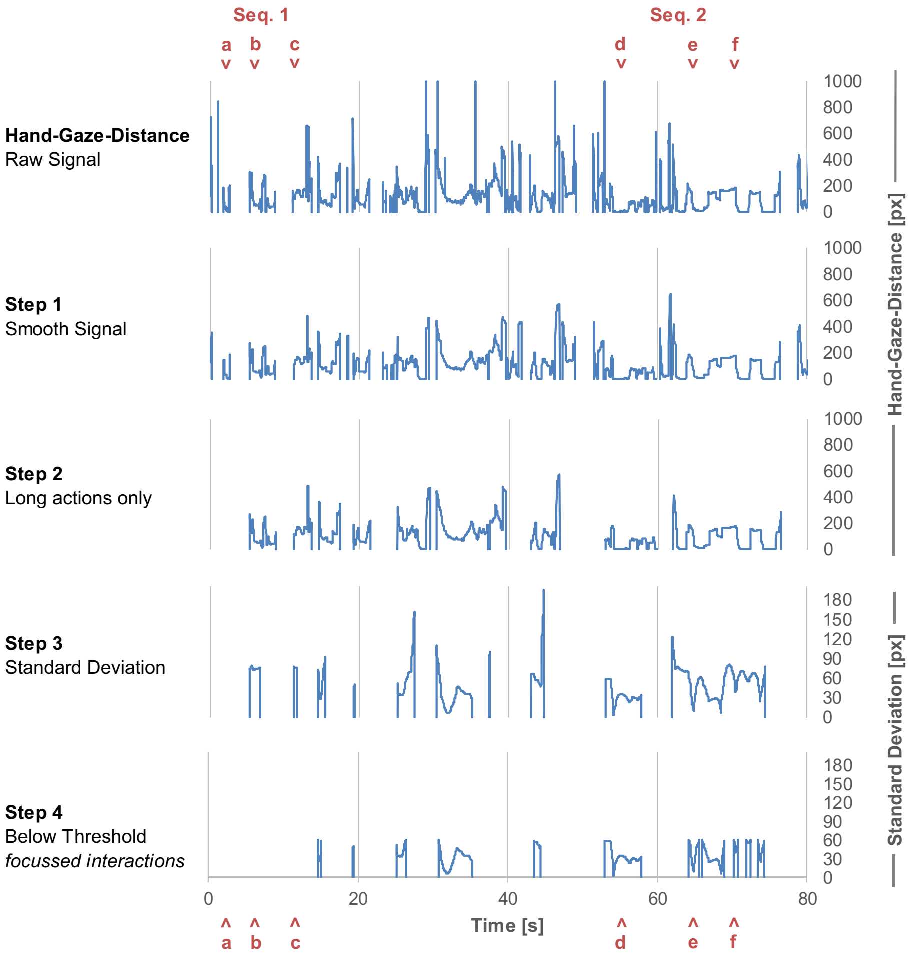

Fig. 2 An excerpt of a participant's data is presented. It takes four steps to isolate the focussed interaction phases from the raw hand-gaze distance. The time stamps *a*–*f* are linked to the images in Fig. 3, showing the corresponding interactions of the same participant

interactions are expected to have a certain minimum duration, in **Step 2** only those sections are kept, in which the hand and gaze were consistently in the image for at least 2 s. In **Step 3**, the standard deviation is calculated within a rolling window of 2 s. This aims to capture also phases in which the HGD is increased due to the usage of a tool. For instance, people using a screwdriver hold it at the handle, but gaze at its tip. The standard deviation considers the constancy of the HGD and thus measures the rigidity of the scene—with and without tool. In **Step 4**, the most rigid interactions are kept by accepting only data below a threshold of 3.5° (60 px). In summary, these four steps isolate periods of long and constant HGD, which are marked as *focussed interactions*.

Deduction of usability problems Video snippets are created automatically for the focussed interactions. They show phases in which the user has a high attentional focus on a performed action. To assess their usability, the snippets are analysed manually and are categorized in problem clusters.

4 Evaluation

The HGD approach was tested on a study of a 3D printer with 40 participants. The participants, all novices on the device, were asked to print a 3D object, to remove the printed object, and to change the filament. The interactions analysed have an average duration of 7.9 min (SD 3.2 min). To exemplify how the algorithm works, two interaction sequences are presented first (Sect. 4.1). Subsequently, the HGD approach is applied to the entire sample. The detection of the hand (Sect. 4.2), the accuracy of the usability problems derived from the focussed interactions (Sect. 4.3) and the manual effort (Sect. 4.4), are evaluated.

4.1 Exemplary sequences

The left and the right column of Fig. 3 show two sequences with similar durations of 10–15 s. The first sequence shows three different actions, namely grasping the scissors (*a*), cutting the filament (*b*) and opening the side door (*c*). All

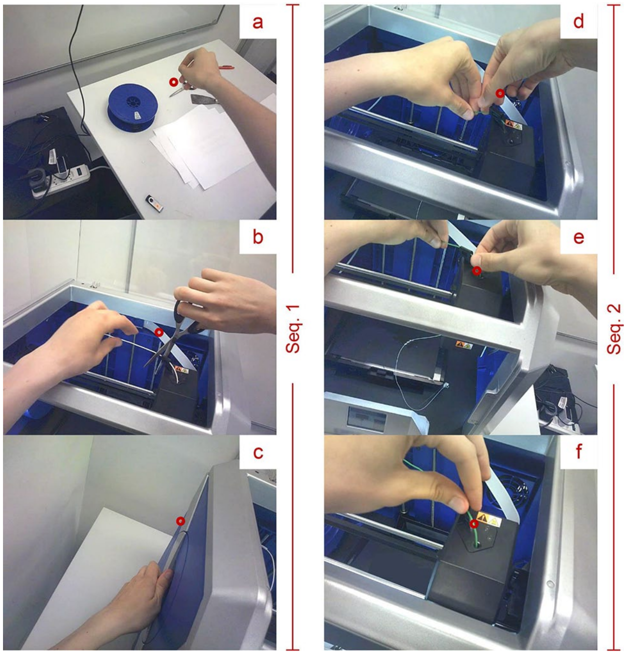

Fig. 3 Two sequences of actions are shown, both in a similar time frame of 10–15 s. The images are linked to the time stamps shown in Fig. 2. Seq. 1 (*a*, *b*, *c*) shows three different actions which are easy to handle, while Seq. 2 (*d*, *e*, *f*) shows one long action that is demanding for the participant

actions are fluid and unproblematic for the user. As visible in Fig. 2, *a* is registered as an action, but is filtered out for not being long enough. The actions *b* and *c* are registered as long actions, but they are filtered out for not being constant enough. Their standard deviations are above the threshold of 60 px.

In the second sequence, all images *d*, *e*, *f* show the same action. The participant encountered difficulties inserting the material into the printing head. This problem, which is also marked in Fig. 4, occurs frequently among the participants. The hands hold the filament while the gaze is rigidly directed to the filament's end for several seconds, which results in a constant HGD, leading to the classification as a focussed interaction.

4.2 Detection of the hands

Method In order to assess the hand detection applied, a manual evaluation has been performed for 35% of randomly selected participants. The phases of the video depicting at least one hand were noted to the tenth of a second. Subsequently, this *manual detection* was compared to the *automatic detection* for each frame.

Results The data of 14 participants with a total video duration of 157 min (564 858 frames) were analysed manually. As shown in Table 1, the colour detection applied identified 77.4% of the frames correctly (c), combining 50.1 and 27.3%. The falsely detected frames consist of 15.6% false positives fp) and 7.0% false negatives (fn).

Table 1 The hand detection is evaluated per frame on the basis of 14 participants, and the ratios of correct (c), false positive (fp) and false negative (fn) frames are reported

	Autom. detection	
	Hand	No hand
Manual detection		
Hand	50.1%^{c}	7.0%^{fn}
No hand	15.6%^{fp}	27.3%^{c}

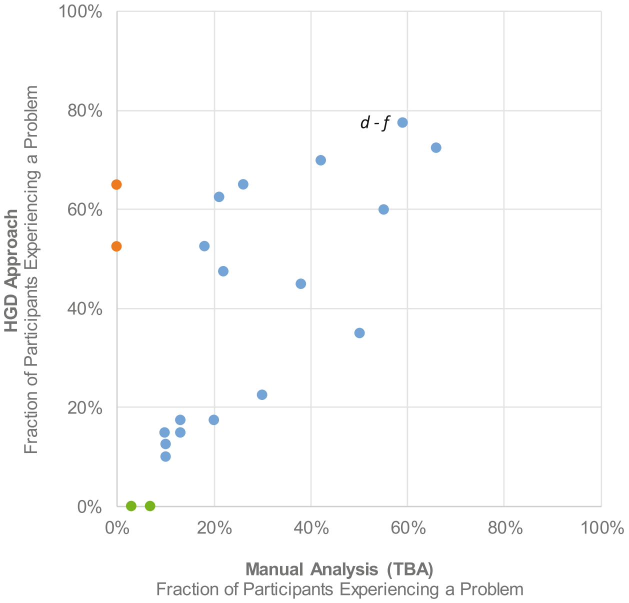

Fig. 4 Each *dot* represents a usability problem. The 17 problems marked with *blue dots* were found with both methods, the *green dots* are problems only found with the manual analysis, while the *orange dots* are problems only detected with the HGD approach. The HGD approach and the manual analysis correlate significantly relating to the fraction of participants experiencing the problems (r = .58, p (two-tailed) <.01.). (Color figure online)

Discussion The frames recorded falsely positive are less problematic for finding usability problems, as they can be eliminated in the subsequent manual analysis. In contrast, falsely negative detected frames are more critical, as they imply a loss of information. In the evaluation presented 7.0% of the frames are false negatives. On the basis of a sample size of 40 participants, it is estimated that this low number has only a weak influence on the evaluation of the HGD approach. In combination with the 77.4% correctly detected frames, the performance of the hand detection is acceptable.

4.3 Identification of usability problems

Method The focussed interactions of all 40 participants were analysed to find out whether they depict a usability problem and if so, they were assigned to problem clusters. Those clusters were then compared to problem clusters derived from a separate manual analysis of the MET videos, performed according to the Target-Based Analysis [14].

Results The HGD algorithm identified 1589 video snippets as focussed interactions across all 40 participants. Of these, 231 (14.2%) were removed since they did not contain a hand.[2] Of the remaining 1358 snippets, 928 (68.3%) show a demanding interaction.

The comparison of the problem clusters with the separate manual analysis showed that each method was able to uncover 19 problems, of which 17 were identical in the two methods. For both methods, the fraction of participants experiencing a certain problem is presented in Fig. 4. Rarely occurring problems are displayed on the bottom left, while frequent problems are shown in the top right position of the diagram. There is a significant positive relationship between the usability problems found with the HGD approach and those detected by a manual analysis, r = .58, p (two-tailed) <.01. For the 17 overlapping problems (blue dots) the HGD algorithm on average detects 11% more participants experiencing a problem than the manual analysis.

In addition to the 17 overlapping problems, the manual analysis uncovered two rarely occurring problems (green dots; 3 and 7% of the participants). These show participants hesitating to start an action. Two problems were exclusively found by the HGD approach (orange dots). Both show users having difficulties with the handling of the filament cartridge.

Discussion The high amount of usability problems in the focussed interactions is a promising result: A total of 68.3% of the focussed interactions with correct hand detection show usability problems. In addition, the approach presented and the manual analysis correlate significantly relating to the fraction of participants experiencing a problem. This result was higher than expected.

Instead of only detecting problems in which the manual interaction is demanding, the HGD approach also found high-level problems of comprehension and confusion. Users who could not find the correct target often resorted to trial and error which again led to demanding interactions, detectable by the HGD algorithm. The example of opening the side door of the 3D printer can illustrate this effect. 45% of the users did not directly realize that the 3D printer has a side door. After a few seconds of hesitating, they tried to remove the filament without opening the side door from inside the printer. As this is not possible, they experienced problems, which resulted in focussed interactions.

However, problems in which the hand is not involved, such as hesitating, cannot be detected with the HGD algorithm. Even though in this study only two problems occurred rarely (3 and 7%), it is a limitation of the approach. The two problems exclusively found with the HGD approach are clearly to be judged as usability problems, but were missed in the manual analysis. They were considered as preparatory steps rather than as part of the actual task. As missed problems due to such subjective restrictions are critical for usability tests, the more objective detection of the algorithm is advantageous in this aspect.

The HGD approach identified usability problems more often than the thorough manual coding, which indicates a conservative setting of the algorithm. To save time in the later stages of the approach, a more aggressive trimming of the parameters should be tested.

4.4 Analysis of video snippets

Method The amount of work required to analyse the video snippets of the focussed interactions is evaluated by comparing the accumulated duration of all the snippets to the total duration of the entire recordings.

Results The total duration of videos to be assessed manually was reduced from 321 min for the entire recordings to 118 min for only the snippets. The snippets have an average duration of 4.4 s (SD 3.7 s).

Discussion The application of the HGD approach reduced the duration of video material to be analysed manually by 63%. As all participants were novices, usually having more difficulties than experienced users, it can be expected that the ratio of focussed interactions is even lower with experienced users.

Watching only the short video snippets with a high probability of containing a usability problem led to a higher sustained concentration of the analyst compared to manual analysis of the entire recording. By presenting only the focussed interactions to the analyst, the subjective nature of deciding on usability problems is reduced. However, it also reduces the understanding of the context. This has been partly compensated by expanding the video snippets for half a second at the beginning and at the end. Furthermore, analysts are advised to watch about three entire recordings before starting to evaluate the snippets in order to gain a high-level understanding of the interaction.

[2] The difference to the ratio of false positives reported in Subsect. 4.2 is due to the filtering described in Sect. 3.

5 Discussion and conclusion

To assess whether cognitive demanding interactions could be automatically detected on the basis of MET videos, a multi-modal feature level fusion considering both the perception (gaze) and the actions (hand) has been performed. The paper presents a first evaluation processing the HGD extracted from MET videos. Long phases of constant HGD are marked as focussed interactions, representing sequences of high attentional focus towards the current action.

To assess the viability of the approach, it was applied to a usability study of a 3D printer with 40 participants. The HGD approach showed similar results and was significantly faster than a manual video analysis. The application of one hand-crafted feature—the HGD—showed acceptable accuracy, within the case at hand, to which the evaluation is limited. However, to broaden and improve the HGD approach we see great potential through the field of AI, in the following three aspects:

1. The colour-based hand detection could be replaced by an object detection algorithm. Considering the varying number of objects (no hand, one hand, two hands) region proposal CNNs could be applied, which would also set the basis for real-time detection [16].
2. Object detection and localization of the product parts, as described by [17], would increase the automation of the HGD approach. Being able to assign the gazes during focussed interactions to parts of the product would allow critical parts to be reported automatically and thus could replace the manual video analysis.
3. Besides the HGD, other features such as pupil size and other patterns such as specific gaze motions might play an important role in detecting demanding interactions. A fully automated end-to-end deep learning approach using CNNs with a multi-modal classifier (video, gaze, pupil size) could be applied. However, learning a global description of the video's temporal evolution, especially for long videos, is considered to be a challenging task [23].

Overall, this investigation shows that studying the interplay between gaze and hand is vital to understand human behaviour. A fluent interaction sequence can be distinguished from demanding handling interactions requiring a high focus of attention. This could be valuable in any scenario where first-person video footage, eye tracking data and handling interactions are available. The automatic identification of demanding sequences is seen as the first step to a real-time event-interpretation of human behaviour on the basis of MET data.

References

1. Bowman M, Johannson R, Flanagan J (2009) Eye–hand coordination in a sequential target contact task. Exp Brain Res 195(2):273–283. doi:10.1007/s00221-009-1781-x
2. Buettner R (2013) Cognitive workload of humans using artificial intelligence systems: towards objective measurement applying eye-tracking technology. In: KI 2013: 36th German conference on artificial intelligence, 16–20 Sept 2013, vol 8077, pp 37–48. Springer, Berlin. doi:10.1007/978-3-642-40942-4_4
3. Crawford J, Medendorp W, Marotta J (2004) Spatial transformations for eye hand coordination. J Neurophysiol 92(1):10–19. doi:10.1152/jn.00117.2004
4. Duchowski A (2007) Eye tracking methodology theory and practice. Springer, London
5. Essig K, Sand N, Schack T, Kunsemoller J, Weigelt M, Ritter H (2010) Fully-automatic annotation of scene videos: establish eye tracking effectively in various industrial applications. In: Proceedings of SICE annual conference 2010, Taipei, pp 3304–3307
6. Hayhoe M, Shrivastava A, Mruczek R, Pelz J (2003) Visual memory and motor planning in a natural task. J Vis 3(1):49–63. doi:10.1167/3.1.6
7. Henderson J (2013) Eye movements. In: Reisberg D (ed) The Oxford handbook of cognitive psychology. Oxford University Press, Oxford, pp 69–82. doi:10.1093/oxfordhb/9780195376746.013.0005
8. Just M, Carpenter P (1980) A theory of reading: from eye fixations to comprehension. Psychol Rev 87(4):329–354. doi:10.1037/0033-295X.87.4.329
9. König P, Wilming N, Kietzmann T, Ossandón J, Onat S, Ehinger B, Gameiro R, Kaspar K (2016) Eye movements as a window to cognitive processes. J Eye Mov Res 9(5):1–16
10. Land M, Mennie N, Rusted J (1999) The roles of vision and eye movements in the control of activities of daily living. Perception 28(11):1311–1328. doi:10.1068/p2935
11. Land M, Tatler B (2009) Looking and acting—vision and eye movements in natural behaviour. Oxford University Press, Oxford. doi:10.1093/acprof:oso/9780198570943.001.0001
12. Li C, Kitani KM (2013) Pixel-level hand detection in egocentric videos. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 3570–3577. doi:10.1109/CVPR.2013.458
13. Mennie N, Hayhoe M, Sullivan B (2007) Look-ahead fixations: anticipatory eye movements in natural tasks. Exp Brain Res 179(3):427–442. doi:10.1007/s00221-006-0804-0
14. Mussgnug M, Sadowska A, Meboldt M (2017) Accepted: Target based analysis—a model to analyse usability tests based on mobile eye tracking recordings. In: Proceedings of the 21st international conference on engineering design (ICED 17), 21–25 Aug 2017. Design Society, Vancouver
15. Mussgnug M, Waldern F, Meboldt M (2015) Mobile eye tracking in usability testing : designers analysing the user–product interaction. In: Proceedings of the 20th international conference on engineering design (ICED 15), 27–30 July 2015. Design Society, Milan, pp 349–358
16. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst: 91–99
17. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) OverFeat: integrated recognition, localization and detection using convolutional networks. https://arxiv.org/abs/1312.6229
18. Song J, Sörös G, Pece F, Fanello S, Izadi S, Keskin C, Hilliges O (2014) In-air gestures around unmodified mobile devices.

In: Proceedings of the 27th annual ACM symposium on user interface software and technology-UIST '14, pp 319–329. doi:10.1145/2642918.2647373

19. Suchan J, Bhatt M (2016) Semantic question-answering with video and eye-tracking data: AI foundations for human visual perception driven cognitive film studies. In: Proceedings of the 25th international joint conference on artificial intelligence (IJCAI-16), 9–15 July 2016. AAAI Press, New York, pp 2633–2639
20. Tatler B, Hayhoe M, Land M, Ballard D (2011) Eye guidance in natural vision: reinterpreting salience. J Vis 11(5):1–23. doi:10.1167/11.5.5
21. Tatler B, Land M (2015) Everyday visual attention. In: Fawcett J, Risko E, Kingstone A (eds) The handbook of attention. MIT Press, Cambridge, pp 391–421
22. Triesch J, Ballard D, Hayhoe M, Sullivan B (2003) What you see is what you need. J Vis 3(1):86–94. doi:10.1167/3.1.9
23. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 7–12 June 2015. IEEE, Boston, pp 4694–4702