ISBN
978-3-031-08647-2

Authors
Cicek, Muratcan
Manduchi, Roberto

Publication Date
2022

DOI
10.1007/978-3-031-08648-9_46

Peer reviewed

Learning a Head-Tracking Pointing Interface

Muratcan Cicek[1] and Roberto Manduchi[1]

University of California, Santa Cruz, USA
mcicek@ucsc.edu

Abstract. For people with poor upper limb mobility or control, interaction with a computer may be facilitated by adaptive and alternative interfaces. Visual head tracking has proven to be a viable pointing interface, which can be used when use of the mouse or trackpad is challenging. We are interested in new mechanisms to map the user's head motion to a pointer location on the screen. Towards this goal, we collected a data set of videos of participants as they were moving their head while following the motion of a marker on the screen. This data set could be used to training a machine learning system for pointing interface. We believe that by learning on real examples, this system may provide a more natural and satisfactory interface than current systems based on pre-defined algorithms.

1 Introduction

Interacting with a computer via traditional interface mechanisms requires good upper limb and hand control. For people with motor impairment, tasks such as typing and pointing (moving the pointer with the mouse or a trackpad) may become difficult or impossible [24]. Assistive technology solutions for computer interaction are designed to receive user input in alternative ways, or through entirely different channels such as sound or vision. For example, dictation systems can be used in lieu of typing, but only if the user has understandable speech (e.g., they may be useless for someone with dysarthria). As for pointing, computer vision-based head tracking has gained popularity in recent years, with several applications using this technology already available [2,6,15,17,18]. These systems use a camera (e.g., embedded in a computer screen) to track the user's head motion using computer vision algorithms. Typically, measurements are taken in terms of a “face box" or of a specific facial figure (e.g. the nose tip [4]). These measurements are then mapped to the pointer location in the screen using a pre-defined algorithm.

A main drawback of this approach is that the mapping from head motion to pointer location is not necessarily representative of the user's intent. For example, moving one's head to the right may lead to a rightwards motion of the pointer that is faster than what the user intended. This may result in an overshoot, which then needs to be corrected by a leftward head motion. In practice, the user needs to learn to use the system with head patterns that may not feel “natural". While these algorithms typically afford some parameter tuning, the general mapping mechanism remains unchanged.

We propose a user-centric approach to designing a pointing algorithm based on head tracking. Rather than imposing a pre-defined algorithm mapping head position to pointer position, we would like to learn a flexible mechanism that adapts to the user's intent. As a first step towards this goal, we created a data set where the measured data (head position from video frames) is associated with the desired location of the pointer. To build such a data set, we resorted to the following strategy. We showed a well-visible marker (a white disk) moving on the screen in specific patterns. While watching the marker moving, participants were asked to move their head "as if" they were controlling the marker themselves. Images were collected by a screen-embedded camera, and time-registered with the location of the marker on the screen at each time. We believe that the videos thus recorded are representative examples of the way participants would move their head if asked to move the cursor to replicate the same trajectories traversed by the pattern they saw moving on the screen.

This paper describe our data collection strategy, which was accomplished remotely due to COVID-19 social distancing constraints. We also present examples of the dynamic of a specific facial feature (the nose tip) while participants were following different trajectories of the pattern in the screen, and provide a simple analysis of the variance of the location measured for this feature across participants.

2 Related Work

There are various head-based pointing methods developed by the researchers including physical head-mounted styluses like Finger-nose [21] for touchscreens, and new sophisticated products like Quha Zono [20] and Glassouse [10]. In addition physical styluses, there are also software-based solutions [3, 6, 15, 17, 18] thanks to the front cameras and the advancements in Computer Vision. Vision-based head-based pointing shares the same fundamental with the gaze-based mechanisms [12,16,19] but instead it tracks the movement of the head rather than eye balls. Comparisons between head-based and gaze-based interactions [1, 13] suggest that head-based techniques are more voluntary, stable and have greater accuracy while gaze-based techniques may be faster for some specific tasks like typing [9].

Most of the visual-based head-based pointing solutions rely on off-the-shelf face tracking algorithms to capture user feedback (i.e. head movement) and convert this input into pointing coordinates on the screen. We know that Enable Viacam [17] benefits from the Haar Cascade algorithm [25] for face detection by evaluating its source code. Camera Mouse [3] also allows users to choose the input mechanism for pointing. Besides face tracking, it also includes point tracking based on simple optical flow calculation. In this setting, users can determine a small patch on their face and the software tries to keep track of this patch on the following frames. On the other hand, HeadGazeLib [5] utilizes the depth sensors of the device the locate user face with respect to the camera. While other methods [4, 15] use advanced deep learning algorithms to detect and track the facial features from RGB images, their conversion functions are again tailored by the developers and involve no machine learning. To the best of our knowledge, there is no visual-based head-based pointing solution that aims to learn pointing from user's appearance directly.

3 Data Collection

We recruited 8 participants (3 female) from our university. One participant has a motor impairment due to cerebral palsy, and is a regular user of head-based pointing technology. Although this is a relatively small sample size, it is adequate for a proof-of-concept. We will consider a larger set of participants (including more participants with motor impairment) in future work. The goal of this study was to collect videos of the participants as they moved their head, following the path of a small white disk shown on their computer screen. The participants were instructed to pretend that they were controlling the white disk with their head motion. They were asked to not just follow the disk with their eye gaze, but by moving their head. No other instructions (e.g., how much to move their head, whether to rotate it vs. move it, etc.) were given. Hence, we can assume that the head motion of each participant was as "natural" and spontaneous as it could be.

We first generated a number of "trajectory videos" with a small white disk moving along a predetermined trajectory against a black background. Some of these trajectories were repeated at a slower velocity. Some trajectories included "pause" points, where the disk would stop for one second. Participants were able to see the future path of the disk (shown with dimmed brightness), so that they would know in advance where the disk would move next. Examples of disk trajectories are shown in Fig. 1. Note that in all trajectories, the disk started and ended at the center of the screen. We uploaded these trajectory videos (17 in total) on YouTube and created separate playlists for each participant, with the order of the video randomly permuted for each playlist.

The study was conducted remotely due to the social distancing requirements imposed by the COVID-19 pandemic. We utilized the Zoom platform to run the data collection sessions, including recording the participants' visual input during the pointing tasks. For each participant, we scheduled a one-hour online meeting via Zoom. We collected information about the computer they would use for the test, whether they would use the embedded camera in the screen or an external camera,the screen size and resolution. In the teleconference, we explained to each participant how the test would be conducted, then asked them to go to the YouTube site at the playlist assigned to them, and to expand the browser window to fulls screen. In this way, participants would only interact with the moving disk in the trajectory videos, while images of their head were taken by the camera and recorded in the cloud via Zoom.

Consecutive trajectory videos within the playlist were separated by 10 second intervals. Participants could use these intervals of time to briefly rest, and they were also allowed to pause the playlist in between trajectory videos. An acoustic signal was played at the beginning and at the end of each trajectory video in the playlist. This was used to synchronize the video displayed to the user, with the video of the user recorded via Zoom (“user video"). These user videos were recorded at a resolution of 1280 × 720 pixels and at a rate of 25 frames per second. The whole session for each participant as recorded by Zoom was exported as a single video for simplicity. We then cropped individual user videos, using as reference the acoustic signal recorded at the beginning and at the end of each trajectory video. In this way, we obtained pairs of synchronized trajectory-user videos, to be used for our analysis. 17 such video pairs were recorded for each user. We had to discard only 2 such video pairs, one due to noticeable latency caused by Zoom, and one because the video was mistakenly interrupted by the experimenter. In total, we obtained 136 synchronized video pairs from 8 separate participants, with the length of the user videos varying between 536 and 2267 frames.

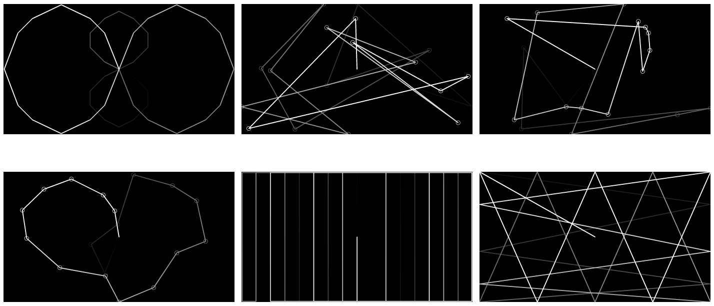

Fig. 1: Samples of trajectory videos. The whole trajectory of the white disk is visible, with lighter color indicating earlier locations in the trajectory. Small circles correspond to location where the white disk stopped for one second.

4 Head Motion Computation

One of the goals of this study is to explore whether the motion of the white disk on the screen could be predicted from the user video. For this purpose, we first extracted a number of visual “features", that can be used to describe the user head’s motion. These features can then be mapped, using suitable machine learning mechanisms, to the position of the disk on the screen.

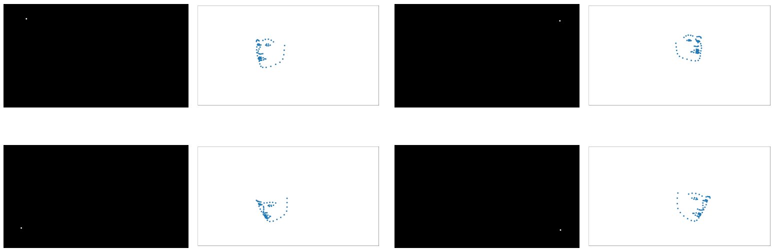

Fig. 2: Facial landmarks produced by the PFLD algorithm [11] for one of the participants, taken at the time the white disk appeared in the location shown in the left half of the figure.

A very simple, though perhaps not very informative, feature is the location of the “face box", defined as a rectangle encompassing the whole face image [7, 14, 23]. A richer description can be obtained by identifying specific facial landmarks. We experimented with three state-of-the-art facial landmark detection models [11, 26, 27]. For example, Fig. 2 shows the location of the facial landmarks produced by the PFLD algorithm [11] for one of the participants, at the times when the white disk being followed was situated in vicinity of the four corners of the screen.

A higher-level feature that we will consider in future work is the pose (3-D location + orientation) of the user’s head, which can be computed using 3-D deformable models (e.g., [8, 22, 28]).

5 Trajectories Analysis

It is instructive to compare the trajectory of the visual features being tracked, against that of the white disk on the screen. This can provide some intuition about how a user would move their head in relation to the desired pointer location. In Fig. 3, we show the trajectory of a specific facial feature, the user’s nose tip, for two participants (P2 and P6), viz-a-viz the trajectory of the white disk. Note that the nose tip location has been used used successfully for head-based pointing control in prior work [4]. While the trajectories of the nose tips may vaguely resemble the trajectory of the white disk on the screen, it is clear that a precise one-to-one positional mapping would be hard if not impossible.

The trajectories of the nose tip feature shown in Fig. 3 are clearly different across the two considered participants. This is to be expected, since the dynamic of head motion associated with tracking the white disk on the screen is completely subjective (remember that participants were not given instructions about how to move their head). In some cases (see e.g. the last case of Fig. 3), a positional bias is visible (possibly because the users positioned themselves at different locations in front of the camera). In these cases, the bias could be easily recovered and compensated for.

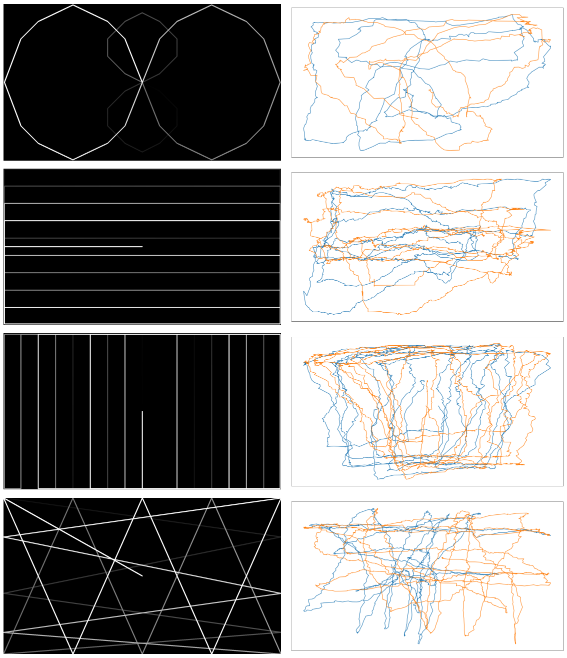

Fig. 3: Trajectories of the nose tip features for two different participants (P2 and P6) associated with the white disk trajectories shown in the left half of each row.

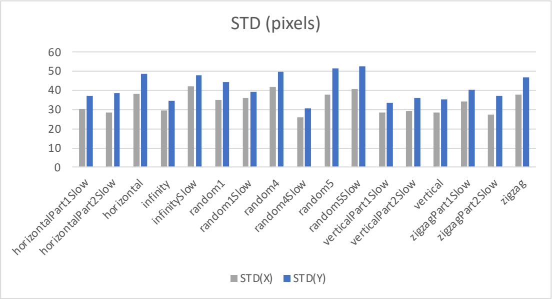

Fig. 4: Average standard deviation of the X and Y coordinate of the nose tip across participants for each trajectory video.

In order to quantify the difference between trajectories across participants, we computed a measure of variance as follows. For each trajectory video, we measured, at each time, the variance in the X and Y coordinate of the nose tip location across all participants. (We excluded P5 in this analysis, as facial feature detection was unreliable for this participant.) Then, we computed the average of these variances over the whole trajectory. The squared root of the average variance (i.e., the standard deviation) for the X and Y coordinates of the nose tip are plotted for each trajectory video in Fig. 4. These values vary between 28 and 42 pixel for X, and between 31 and 53 pixels for Y (remember that the recorded images have resolution of 1280 × 720 pixels.)

6 Conclusions

This paper presents a unique data set collected for the purpose of understanding the different head motion dynamics adopted by different participants while imagining to control a moving disk on a screen. We are currently using this data set to train a machine learning system that can predict the desired location of the cursor based on the user's head motion. Our hope is that, by learning from videos collected in response to a stimulus on the screen, this system can do a better job of mapping image feature to cursor locations than current, hand-tailored algorithms.

Our initial analysis of the collected data shows that there is a fairly large variance of the location of facial features (e.g., the nose tip) across participants while following the same disk trajectory. This suggests that a certain degree of personalization may be necessary, in order to adjust the algorithm to the specific head dynamics of each user.

References

1. Bates, R., Istance, H.O.: Why are eye mice unpopular? a detailed comparison of head and eye controlled assistive technology pointing devices. Universal Access in the Information Society **2**(3), 280–290 (2003)
2. Betke, M., Gips, J., Fleming, P.: The camera mouse: visual tracking of body features to provide computer access for people with severe disabilities. IEEE Transactions on neural systems and Rehabilitation Engineering **10**(1), 1–10 (2002)
3. of Boston College, T.: Cameramouse (2018), retrieved January 31, 2022 from http://www.cameramouse.org/
4. Cicek, M., Dave, A., Feng, W., Huang, M.X., Haines, J.K., Nichols, J.: Designing and evaluating head-based pointing on smartphones for people with motor impairments. In: The 22nd International ACM SIGACCESS Conference on Computers and Accessibility. ASSETS '20 (2020)
5. Cicek, M., Xie, J., Wang, Q., Piramuthu, R.: Mobile head tracking for ecommerce and beyond. Electronic Imaging **2020**(3), 303–1 (2020)
6. Corporation, O.I.: Headmouse nano (2017), retrieved January 31, 2022 from http://www.orin.com/access/headmouse/
7. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. CoRR (2019), http://arxiv.org/abs/1905.00641
8. Fanelli, G., Gall, J., Van Gool, L.: Real time head pose estimation with random regression forests. In: CVPR 2011. pp. 617–624. IEEE (2011)
9. Gizatdinova, Y., Špakov, O., Surakka, V.: Comparison of video-based pointing and selection techniques for hands-free text entry. In: Proceedings of the ACM international working conference on advanced visual interfaces. pp. 132–139 (2012)

10. Glassouse: Glassouse assistive device (2018), retrieved July 17, 2018 from http://glassouse.com/
11. Guo, X., Li, S., Zhang, J., Ma, J., Ma, L., Liu, W., Ling, H.: PFLD: A practical facial landmark detector. CoRR (2019), http://arxiv.org/abs/1902.10859
12. Inc., A.: Use switch control to navigate your iphone, ipad, or ipod touch (2018), retrieved July 15, 2018 from https://support.apple.com/en-us/ht201370
13. Kytö, M., Ens, B., Piumsomboon, T., Lee, G.A., Billinghurst, M.: Pinpointing: Precise head-and eye-based target selection for augmented reality. In: Proceedings of the 2018 ACM CHI Conference on Human Factors in Computing Systems. p. 81 (2018)
14. Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., Huang, F.: DSFD: dual shot face detector. CoRR (2018), http://arxiv.org/abs/1810.10220
15. LLC, P.D.: Smylemouse (2016), retrieved January 31, 2022 from https://smylemouse.com/
16. Majaranta, P.: Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies: Advances in Assistive Technologies. IGI Global (2011). https://doi.org/10.4018/978-1-61350-098-9, https://doi.org/10.4018/978-1-61350-098-9
17. Mauri, C.: Enable viacam (2017), retrieved January 31, 2022 from http://eviacam.crea-si.com/index.php
18. Mauri, C.: Eva facial mouse (2018), retrieved January 31, 2022 from https://github.com/cmauri/eva_facial_mouse
19. MyGaze: Mygaze assistive (2018), retrieved July 16, 2018 from http://www.mygaze.com/products/mygaze-assistive/
20. oy., Q.: Quha zono (2018), retrieved July 15, 2018 from http://www.quha.com/products-2/zono/
21. Polacek, O., Grill, T., Tscheligi, M.: Nosetapping: what else can you do with your nose? In: Proceedings of the 12th ACM International Conference on Mobile and Ubiquitous Multimedia (2013)
22. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 2074–2083 (2018)
23. Tang, X., Du, D.K., He, Z., Liu, J.: Pyramidbox: A context-assisted single shot face detector. CoRR (2018), http://arxiv.org/abs/1803.07737
24. Turturici, M., Fanucci, L.: Inertial human interface device for smartphone and tablet dedicated to people with motor disability. In: Volume 33: Assistive Technology: From Research to Practice. Assistive Technology Research Series (2013). https://doi.org/10.3233/978-1-61499-304-9-494
25. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. vol. 1, pp. I–I. Ieee (2001)
26. Wang, X., Bo, L., Li, F.: Adaptive wing loss for robust face alignment via heatmap regression. CoRR (2019), http://arxiv.org/abs/1904.07399
27. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. CoRR (2018), http://arxiv.org/abs/1805.10483
28. Yang, T.Y., Chen, Y.T., Lin, Y.Y., Chuang, Y.Y.: Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1087–1096 (2019)