Investigating Webcam-based Hand-tracking for Navigation in Micro-task Crowdsourcing

Safouane El Hilali
Supervisor(s): Garrett Allen, Dr. Ujwal Gadiraju
EEMCS, Delft University of Technology, The Netherlands
20-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfillment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

The health of micro-task crowdsourcing workers, also called crowdworkers, is something that is overlooked in the micro-task crowdsourcing literature. Due to repetitive tasks, they can develop Repetitive Strain Injuries. To look into other ways of navigating Crowdsourcing Work Environments (CSWEs) outside the mouse and keyboard paradigm, we consider webcam-based hand-tracking in this paper. The main question we considered was which hand gestures were most suitable for navigating CSWEs. By having micro-task crowdworkers (n=14) test five methods of navigating CSWEs, we found that gestures which were considered easiest and most useful were those that specified a single action in an interface catered to hand-tracking controls. Gestures which attempt to directly replace the mouse in a regular mouse-oriented interface were rated lower on usefulness and ease of use. We also found that most crowdworkers were unlikely to use hand gestures for progressing through related subtasks, since they were considered harder than using the keyboard and mouse.

1 Introduction

Micro-task crowdsourcing is a type of work that is completed by individuals who are paid very small amounts of money for each task that they complete. The tasks are usually simple and quick to do, and they can be done online from anywhere in the world. This type of work has become very popular in recent years, as it allows businesses to get small tasks done quickly and for cheap. There are a number of different platforms (e.g. Amazon MTurk[1], Microworkers[2]) that allow businesses to post tasks, and micro-task solvers (crowdworkers) to find and complete them.

However, there is little research on the health of crowdworkers, but research done on office workers in similar environments has shown that musculoskeletal complaints as a result of repetitive movement are common [6]. Signs of injury in the neck, wrist, shoulders and general upper-body are called Repetitive Strain Injuries (RSI) [2]. We can deduce that crowdworkers are even more at risk of RSI than office workers, because micro-tasks in particular involve a lot of repetitive and monotonous actions performed with mouse and keyboard [4].

Taking a break from using the keyboard and mouse helps prevent RSI [5]. Hence, we are looking at real-time hand-tracking, which is another modality for interacting with the PC, and navigating through human-interface tasks (HITs) and their sub-HITs, which are commonly performed actions on crowdsourcing work environments (CSWEs). So this study aims to find out which hand gestures are most suitable for allowing crowdworkers to use real-time hand-tracking in order to navigate their CSWE.

To do this, we aim to answer the following questions:

- **Research Question**: What hand gestures are best suited for navigation in micro-task crowdsourcing?
- **Subquestion 1**: How can hand gestures be classified?
- **Subquestion 2**: What actions are commonly performed while navigating through HITs and sub-HITs?

In the section 2, we go over the theory needed to answer the first subquestions and the way the final research question will be answered with a user-study. Section 3 describes the methods used to conduct the research. Section 4 presents the results of the user-study. Section 5 discusses the ethical and responsible conduct of the research. Finally, sections 6 and 7 discuss the implications of the experiment and provide a conclusion.

2 Background

2.1 Theory

Hand-tracking

Since the word "gesture" can mean a lot of things, it's important to clarify what is meant in this paper when we use the word "gesture." The word is derived from the Latin word gestus, meaning "action, movement." Since we only concern ourselves with hand gestures, Vuletic et al. [10] have provided a definition of hand gestures in their human-computer interaction (HCI) paper, as "gestures performed using one or both hands, including finger gestures when they were performed along with a number of other varied gestures e.g. pointing gesture is used for selection of an object and then pinching gesture is used to deform that object, or to move it to a different location."

As the first subquestion asks what types of hand gestures exist, we need a categorization or taxonomy of some kind to bring order to the set of all possible gestures.

The literature has many differing taxonomies that have been proposed to classify gestures. It is widely accepted that gestures can be either *communicative (social)* or *manipulative (functional)* [3] [8]. Communicative gestures are meant to communicate meaning between two actors (e.g. pointing at an object to draw another person's attention to it, or flapping two hands to symbolize a bird). In regular settings, they often accompany speech. Manipulative gestures instead are meant to act on "objects in an environment" [8]. Vuletic et al.'s [10] definition therefore pertains to manipulative gestures, and all mention of hand gestures in this paper hereafter will consider that subset of gestures.

The literature on functional gesture recognition is large and diverse. Carfi and Mastrogiovanni [3] have reviewed the literature of gesture taxonomies up until 2022, and have proposed a new design in which manipulative gestures are described by four features, *effect*, *time*, *focus* and *space*.

1. *Effect* refers to how a gesture is going to affect the machine the human is interacting with. This is further divided into two categories, *continuous* and *discrete*. A continuous gesture keeps interacting with the system for as long as it is active, e.g. the cursor moving the same directions as the hand. A discrete gesture is one where the whole gesture performs a single input, like pinching in order to press the Enter key.

[1] https://mturk.com

[2] https://microworkers.com

2. *Time* divides gestures into the classes of *dynamic* and *static*. A dynamic gesture is one that includes multiple poses over time, like waving, while a static gesture is one that only includes one static pose, like a thumbs up.
3. *Focus* refers to which body part is relevant for the gesture, which does not divide gestures in classes but just describes each gesture using the name(s) of the relevant body part(s).
4. *Space* determines whether the meaning associated with a gesture depends on the physical location where you perform it. Tapping virtual buttons in the air makes the gesture of tapping have different purposes depending on the space it is performed in.

Together, these four features make it possible to describe every gesture uniquely.

This study will only consider hand gestures which can be tracked using a PC or laptop's webcam. There are several consumer products on the market that allow users to track their hand and fingers with great accuracy, like the Leap Motion Controller [9]. These gadgets are not considered in this study and the focus is only on usage of regular webcams. This is because crowdworkers in a study by Deng and Joshi have revealed that one of the biggest factors in their participation in micro-task crowdsourcing is equipment affordability [4]. 99.5% and 98.9% of respondents from the United States and India respectively reported they use a laptop or PC to work on micro-tasks, in a study by Newlands and Lutz [7]. Since many PCs and almost all laptops are equipped with webcams, making use of this input as an alternative input modality would allow the largest amount of micro-task crowdworkers to benefit.

CSWE Navigation

To answer the second subquestion, we took a look at the CSWEs Prolific[3] and Amazon MTurk[4]. As one can see in Figure 1, upon opening the Amazon MTurk work environment, the crowdworker sees a list of *HIT groups* they can participate in, which refers to a group of similar tasks requested by one requester. The information displayed for each HIT group is the name of the requester, the title of the group, the payment that workers get, and a button that either lets the worker accept and start working on the HITs, or to apply for qualification. On Prolific, the interface also shows a list with titles, requesters and compensations, but the compensation shown is hourly.

Instead of clicking accept and immediately starting to work on a HIT group, the crowdworker could also go through the entire list of HIT groups and add those that are pending qualification or accepted to a backlog of tasks that they can go back to later. This way they can spend some time building a batch of backlogs to go back to later.

To navigate through HITs, crowdworkers need to be able to move forward, and if the requester allows it, to return to the previous HIT in order to correct mistakes.

In short, to navigate through a CSWE environment, a crowdworkers needs to be able to see all available tasks and accept or reject them, and while working on subtasks, they need to be able to move forwards and backwards between them. These actions capture the essential actions that are needed to navigate through a CSWE environment.

[3] `https://app.prolific.co`

[4] `https://mturk.com`

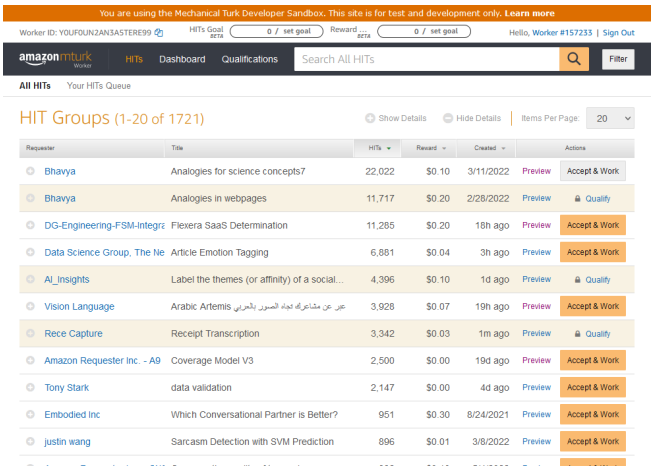


Figure 1: The worker sandbox of Amazon MTurk

2.2 Practice

With the taxonomy of gestures established, we have decided on what classes of gestures to test by making a division so that we have the most contrasting types of gestures possible. We get these by permuting the first two features of the previously mentioned taxonomy; effect and time. By doing this, we get three classes of gestures: *continuous dynamic gestures*, *discrete dynamic gestures*, and *discrete static gestures*. The taxonomy excludes the possibility of continuous static gestures, since a static gesture cannot continuously influence the system state

To measure the *suitability* of these three classes of gestures, crowdworkers have watched a video where they followed along with instructions to simulate these three types of gestures to accept or reject a list of available tasks, and two types of gestures to navigate between subtasks, as if they were using hand-tracking controls through a webcam.

After simulating these gestures through the follow-along video, the workers gave their opinions on the usefulness and user-friendliness of these gestural controls on open-ended questions and a Likert scale, and finally gave their general opinion on webcam-based hand-tracking for navigation through another open-ended question.

2.3 Contribution

By eliciting the opinions of crowdworkers on webcam-based hand-tracking for navigating CSWE environments, we hope to give CSWE developers and requesters on those platforms more insight in what alternative modalities to provide for crowdworkers.

3 Methodology

As explained in §2.2, a survey and accompanying 'follow-along' video were constructed and sent out to crowdworkers

on a CSWE. The video which the participants watched and followed along with can be viewed on YouTube.[5]

Participants. The participants who participated in this experiment and contributed their feedback came from the crowdsourcing work environment Prolific. The number of participants was 14. Their ages were diverse, with 50% between 18 and 35 years old, 36% between 35 and 45 years old, and 14% being older than 45. Most of the participants (9) have only worked on CSWEs for less than two years, four others have only worked between two and four years, and one between four and eight years. The majority of them (12) have never worked on CSWEs other than Prolific.

Procedure. The tasks which the participants solved consisted of two parts, with three and two different types of interactions per part. The first part consisted of trying out three different ways of browsing the list of HITs using hand-gestural interaction, so that the participants could physically try out the different types of movements. These consisted of

1. *A combination of continuous dynamic and discrete dynamic hand gestures.* In this task, the participant used the continuous dynamic gesture of pinching and moving his hand in order to simulate moving the mouse and scrolling up and down, and the discrete dynamic gesture of pinching the fingers in order to simulate a mouse click. Using these controls, the participant was to scroll a web page modeled after Amazon MTurk's HITs interface and accept marked tasks.
2. *Discrete dynamic hand gestures.* In the second task, the participant saw the title of a HIT, its description, pay, and rating, and could accept this HIT by *swiping* their hand in the air towards the left in order to reject or to the right in order to accept the card.
3. *Discrete static hand gestures.* This task is similar to task two, with the only difference being that the participant did not swipe the card towards the left or right, but held their hand in a *thumbs up* or *thumbs down* pose in order to accept or reject the task, respectively.

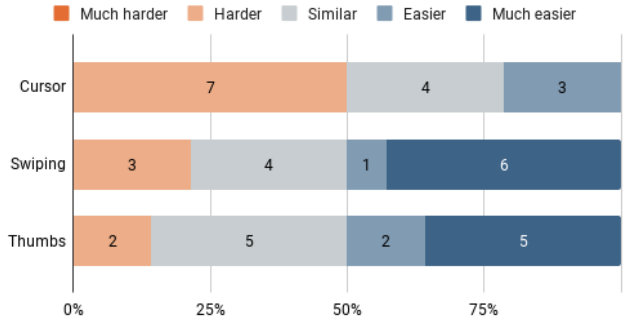
The second part consisted of two different ways to navigate through subtasks in a HIT group. The participant emulated selecting an answer with their keyboard and then using the following types of gestures to move to the next HIT.

1. *Discrete dynamic hand gestures.* The participant simulated pressing a button to label an image as containing a cat or a bird, then swiped in the air to the right, to symbolize the current subtask moving to the left while the next subtask came in from the right.
2. *Discrete static hand gestures.* The participant simulated pressing a button to label an image as containing a cat or a bird, and then held their hand up and pointed to the right to move to the next subtask.

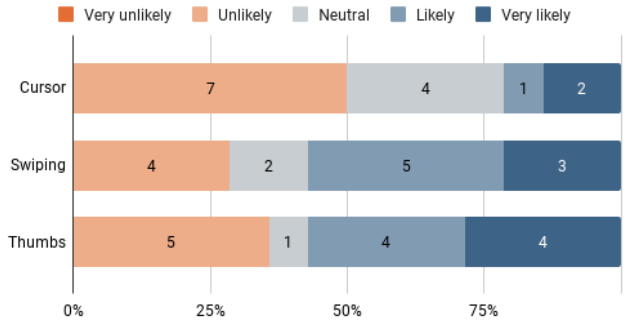
This part did not include a *continuous hand gesture*, since it is not feasible for a worker to interact with the CSWE environment continuously to navigate between subtasks, as navigation is only intended to move to the next or previous subtask, which is a discrete action.

[5] `https://youtu.be/3pYPx2AsAN0`

After each of the two parts, the participants were asked two questions about the hand-tracking methods they just completed: *"Compared to choosing tasks with keyboard-and-mouse controls, how easy were those methods to use?"*, and *"If you could use these hand-tracking controls along with mouse-and-keyboard controls, how likely are you to use them?"*. They motivated their response to the second question in text. Finally, they were invited to give their opinions on webcam-based hand-tracking for navigation in general.



(a) Responses to the question: *Compared to choosing tasks with keyboard-and-mouse controls, how easy were those methods to use?*, concerning the first three methods.



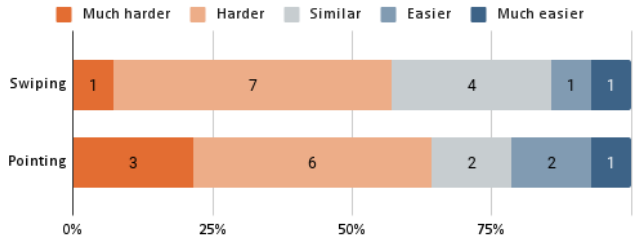
(b) Responses to the question: *If you could use these hand-tracking controls along with mouse-and-keyboard controls, how likely are you to use them?*, concerning the first three methods.

Figure 2: Responses to the Likert scale questions of the questionnaire, concerning the first three methods.

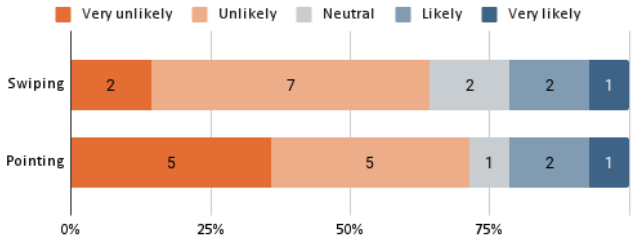
4 Results

The answers to the Likert scale questions regarding ease of use and usefulness of all five methods are displayed in Figures 2 and 3.

As shown in Figure 2a, the participants did not find the hand-tracked cursor method easy to use, seeing as 50% of participants found it harder to use than keyboard-and-mouse (K&M) controls. The swiping and thumbs up/down methods fared better, with only three participants finding swiping harder to use than K&M controls and two participants finding the thumbs up and down harder. Half of the participants found it easier or much easier to pick out tasks with these controls.



(a) Responses to the question: *Compared to choosing tasks with keyboard-and-mouse controls, how easy were those methods to use?*, concerning the last two methods.



(b) Responses to the question: *If you could use these hand-tracking controls along with mouse-and-keyboard controls, how likely are you to use them?*, concerning the last two methods.

Figure 3: Responses to the Likert scale questions of the questionnaire, concerning the last two methods.

Half of the participants reported that they were unlikely to use the first method (Figure 2b) if they were available as an alternative to K&M controls. On the other hand, a majority reported that they were likely or very likely to use the second and third methods if given the option.

As for why they were likely or unlikely to use these methods, the answers were varied. Three participants were worried that inaccuracy could lead to missing work opportunities; one of them responded with *"I think for the swiping in particular you could do the wrong movement"*. Three participants mentioned unintuitiveness of the gesture controls or existing familiarity with K&M controls as a reason. Six participants reported that they found (some of) these methods easy to use, and were therefore likely to use them if given the option.

The hand-tracking methods for progressing through subtasks were less popular with the participants. Over half of the participants reported that those two methods were harder or much harder to use than K&M controls (3a). 64% and 71% of the participants mentioned they were unlikely or very unlikely to use the swiping and pointing method respectively if given the option (Figure 3b.

As motivation, six participants mentioned these methods as being hard: *"It feels like more work because I am less used to the motions"*, *"It felt like there was too much going on"*, *"It's far easier for user to just press 1 or 2 for bird/cat."*. Two participants were not likely to use these methods because they were confusing. One person identified the pointing gesture as being a strained repetitive movement, so not willing to use that. Two participants mentioned accuracy as a point of worry, but were willing to try it out. Two participants reported being willing to use these methods, one mentioned the reason as being fun and easy.

As for their overall thoughts, three participants found hand-tracking controls wholly unnecessary. They found that it overcomplicates things and would rather use a mouse and keyboard. Similarly, one participant found that it might benefit some people, but also slow down too much of the process. Two participants mentioned accuracy, with one of them needing assurance of high accuracy before using it. Four participants were willing to try it out. One participant reported not liking "the idea of having to have a webcam always on," which means privacy concerns may play a role with this modality.

5 Responsible Research

Since this research required human participants, in accordance with good research ethics, we sought approval for human experimentation from the Human Research Ethics Committee (HREC).

Part of good ethics is also to make sure human participants are well compensated for their time. All participants were paid £8.55/hr, the recommended amount by Prolific, and were paid their dues within two hours of completing the study.

All except two of the received answers were included in the study. These two were rejected because the time taken for the study was only around three minutes and 30 seconds. It is not possible to follow along with the video and answer the questionnaire within such a short timeframe. This fact combined with the low quality of their answers indicated that these two workers did not do the study in good faith.

Apart from these exceptions, the responses of all 14 other participants were accepted and incorporated into the results, even if they were hard to interpret or gave only superficial feedback like *"Yes, I'd use it."*. Disappointing or counter-expectational responses were also accepted, since the purpose of this study is not to propagandize for new input modalities but to get the honest opinions of micro-task crowdworkers on them.

6 Discussion

From getting experience with doing research on Prolific, we found that accepting tasks and working on them in batches might not be ideal. Seeing how almost all responses were filled a mere ten minutes after releasing the study, we saw that working on tasks right away is very important. Once the crowdworker is done with accepting and rejecting the available tasks, some might not even be available anymore once the worker wants to start working on them.

One good point mentioned by a participant is that they *"might as well use my mouse and keyboard rather than taking my hands on and off."* It would not make a lot of sense for a crowdworker to use their keyboard and mouse to solve a subtask and their webcam to move to the next one. Raising and lowering their hand would make them expend more energy than simply pressing a button next to their finger to continue. For other researchers looking into alternative modalities for micro-task crowdsourcing, we would recommend not designing experiments where the participants constantly need

to switch between K&M controls and the other modality, but instead have as much of the controls as possible be covered by one modality.

As the discrete hand gestures were most positively received by participants, we recommend other researchers to focus on more testing of discrete hand gestures. Not just for micro-task crowdsourcing, but other kinds of computer users as well.

Due to budgetary constraints, only 14 participants could partake in this study. This makes the results not that statistically significant. Yet we received varied responses, giving a range of supportive and critical opinions. This might well be a representative sample, and the trend we saw of participants to deem the last two methods less useful and harder to use than the first three methods will probably still hold for larger samples. However, the finer details of what methods specifically are more liked by crowdworkers can significantly change with a larger sample size.

6.1 Limitations

This subsection describes a couple of limitations of this study.

Most of the crowdworkers have only done micro-task crowdsourcing on Prolific. Just two of them have used other platforms. On Prolific, workers participate in "studies" instead of "HITs", which means the participants do not have much experience with the truly repetitive and straining tasks that are published on platforms like Amazon MTurk, like data classification tasks.

14 is a small sample size. Ideally, there should be more participants in a future study. A future study should also implement a real interactive web application that the participants use, instead of emulating one by following along with a video. A web application provides more feedback and ensures the participants are actually doing the experiments as they should.

7 Conclusion and future work

The main research question was what hand gestures are best suited for navigation in micro-task crowdsourcing. To this end, we looked into how hand gestures can be classified and landed on the four categories of effect, time, focus, and space. The navigational actions that crowdworkers need to perform on CSWEs are browsing their available tasks and accepting or rejecting them, and moving forwards and backwards through subtasks.

Through an experiment on the CSWE Prolific, we found that crowdworkers were likely to use discrete dynamic and discrete static hand gestures in order to accept or reject tasks. However, this could likely not be implemented on all CSWEs as some, like Prolific, have tasks that are time-sensitive and should be solved as soon as they are accepted.

The crowdworkers were less enthusiastic about using hand gestures for progressing through subtasks, since this was hard and confusing compared to K&M controls. It is therefore not recommended for requesters to implement an option for hand-tracking controls just to navigate through subtasks. Future research should go into a combination of hand-tracking for navigation and for solving the tasks themselves, so that crowdworkers use the keyboard and mouse only minimally. The work of Ajandisz [1] can be used as a starting point for this.

Future studies into this area should also preferably have a larger sample size than 14, and get test subjects from a CSWE that are known for worse working conditions than Prolific.

References

[1] Andris Ajandisz. Investigating gestures of the body as means of input modalities in crowdsourced microtasks.

[2] PM Bongers, HC de Vet, and BM Blatter. Repetitive strain injury (rsi): occurrence, etiology, therapy and prevention. *Nederlands tijdschrift voor geneeskunde*, 146(42):1971–1976, 2002.

[3] Alessandro Carfì and Fulvio Mastrogiovanni. Gesture-based human-machine interaction: Taxonomy, problem definition, and analysis. *IEEE Transactions on Cybernetics*, 2021.

[4] Xuefei Nancy Deng and Kshiti D Joshi. Why individuals participate in micro-task crowdsourcing work environment: Revealing crowdworkers' perceptions. *Journal of the Association for Information Systems*, 17(10):3, 2016.

[5] Robert A Henning, Eric A Callaghan, Anna M Ortega, George V Kissel, Jason I Guttman, and Heather A Braun. Continuous feedback to promote self-management of rest breaks during computer use. *International Journal of Industrial Ergonomics*, 18(1):71–82, 1996.

[6] Alwin Luttmann, Klaus-Helmut Schmidt, and Matthias Jäger. Working conditions, muscular activity and complaints of office workers. *International Journal of Industrial Ergonomics*, 40(5):549–559, 2010.

[7] Gemma Newlands and Christoph Lutz. Crowdwork and the mobile underclass: Barriers to participation in india and the united states. *new media & society*, 23(6):1341–1361, 2021.

[8] Vladimir I Pavlovic, Rajeev Sharma, and Thomas S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):677–695, 1997.

[9] Ultraleap. Tracking — Leap Motion Controller — Ultraleap — ultraleap.com. https://www.ultraleap.com/product/leap-motion-controller/. [Accessed 30-May-2022].

[10] Tijana Vuletic, Alex Duffy, Laura Hay, Chris McTeague, Gerard Campbell, and Madeleine Grealy. Systematic literature review of hand gestures used in human computer interaction interfaces. *International Journal of Human-Computer Studies*, 129:74–94, 2019.