Obtaining Labels for In-the-Wild Studies: Using Visual Cues and Recall

Andrew Vargo, *Osaka Prefecture University, Sakai, 599-8531, Japan*
Shoya Ishimaru, *University of Kaiserslautern, Kaiserslautern , 67663, Germany*
Md. Rabiul Islam, *Osaka Prefecture University, Sakai, 599-8531, Japan*
Benjamin Tag, *University of Melbourne, Parkville, VIC, 3010, Australia*
Koichi Kise, *Osaka Prefecture University, Sakai, 599-8531, Japan*

The observer effect found in laboratory studies has long posed a problem for researchers. In-the-wild studies reduce the observer effect, but have problems with gathering accurately labeled data usable for training algorithms. Manual labeling is time-consuming, obtrusive, and unfeasible, and if done by the researchers, it potentially violates the privacy of the participants. In this article, we present a labeling workflow based on an in-the-wild study that investigated cognitive state changes through eye-gaze in naturalistic settings. We contribute a setup that enables participants to label their data unobtrusively and quickly. We use J!NS MEME electrooculography glasses, Narrative Clip 2 wearable cameras, and a proprietary data tagging software package. Our setup is reproducible for field studies, preserves data integrity, and maintains participant privacy. This workflow can be extended to other studies in pervasive and ubiquitous computing and is especially suitable for deployment in the pandemic and postpandemic world.

In-the-wild studies make it possible for us to accumulate naturalistic data enabling us to validate the applicability of applications and platforms in everyday situations.[1] However, not only can in-the-wild studies be expensive and complex to implement, it is also difficult to acquire accurate data labels that are necessary to train machine learning algorithms without interrupting the participants' lives or violating their privacy.[2] In *laboratory* studies, it is easier to gather labels that researchers can validate as being accurate, but the data collected may not replicate natural behaviors due to tight controls and an artificial environment. The result does not present an easy path toward realistic and accurate data labels to describe naturalistic behavior. Therefore, we need to find ways to allow study participants to successfully, quickly, and privately label their own data in-the-wild.

Natural behavior is an enigma; we know it exists, but it is difficult to approximate it to a satisfactory level. Natural behavior can be defined as being "behavior without boundaries" where individuals act with no restrictions. Obtaining data that perfectly reflect natural behavior, at least with physiological sensing, is (still) unachievable. Ethics, human rights, privacy, and informed consent mean necessitating a high-level of interaction with the research subjects. Even without these restrictions, unless there is a *panopticon* platform, there is a limited ability to have all the essential information to reliably create accurate labels.

The challenge and importance for researchers to obtain accurate *labels* cannot be underestimated. We can imagine a scenario in which researchers want to measure a physiological response to certain everyday activities. For instance, if we want to measure fatigue when users engage with their smartphones throughout the day, we need to differentiate and label when and which activities are occurring. Labels provide the essential training data for machine learning technologies.

Digital Object Identifier 10.1109/MPRV.2021.3129500
Date of publication 17 December 2021; date of current version 11 March 2022.

Deep learning algorithms that classify behaviors automatically are only accurate if they have sufficient training data. A lack of data or properly labeled data can render the technology unusable, but labeling for human activities is often tedious and difficult.[3] This shows the need to develop tools and methodologies for data labeling in numerous domains.

This article is motivated by a practical and specific research question; How can we detect and classify an everyday activity (here: reading) in naturalistic settings for the purpose of building systems that support learning and focus? The benefits of reading and increased reading volume have been studied to great lengths in education and cognitive sciences.[4] Reading can increase the size and retention of vocabulary, and expand abstract reasoning skills. Our studies focus on defining the reading habits of people in everyday settings by deploying ubiquitous tools, thus making it possible to build applications and supplements that can help individuals increase their desired reading habits. To do this, we need training data that accurately describe real daily life.

In this article, we contribute a workflow to accurately label data in-the-wild. In addition, we present a field study investigating reading activities, where we successfully applied this workflow. We outline a methodology that mitigates the Hawthorne effect and maintains participant privacy by using software tools that allow the participants to easily label their data. The methodology that is presented is especially applicable for small-scale studies that focus on specific application areas and need to be repeated or run in iterative cycles. We present a study in which participants were asked to engage in reading activities while wearing electrooculography glasses and a device that records pictures of the participants' point of view throughout their day. We introduce two tools that are customized to work with their relative technologies, but have a logical workflow in many different situations. The main finding of our study is that participant data labeling is effective when the labels are clearly defined and participants engage in the labeling the same day the data are recorded. Deriving from this, we provide insights into what guidelines must be given to the participants in this kind of out-of-the lab study.

BACKGROUND

How do we get usable labels for reading detection? Laboratory studies are good due to the amount of control researchers have over potential confounding influences. Usage of devices can be easily managed and settings can be controlled.[5] Most importantly, researchers can be confident that the labels they are collecting are accurate through observation or preprocessing of tasks. Using the example of fatigue and smartphone usage, researchers could easily label the usage through direct observation. However, there are drawbacks. The artificial nature of the laboratory setting may make participants act unnaturally and the direct oversight from researchers can create a strong observer effect. As made famous in the Hawthorne Works studies during the 1920s and 1930s, the *Hawthorne* effect describes a situation in which study participants modify their behavior due to the knowledge that they are being observed.[6] This can make the labels unsuitable for classifying naturalistic behavior. In some extreme cases, the results from laboratory studies will differ greatly from those done outside.[7] In addition, it becomes prohibitive to engage in the experiment for any significant length of time if it is difficult to bring participants into laboratories.

In-the-wild studies are the obvious choice of methodology for many research projects, including this one. They allow participants to provide records of behaviors in the course of their daily lives, which should produce more natural labels. Also, the experiment can be run for days or weeks at a time. However, in-the-wild studies are not free from observer effects. Therefore, it is beneficial for researchers to make their data collection process clear to the participants and reduce the amount of direct observations on the participant data.

The need for accurate labels for training data requires that we consider how we label the recorded data. One obvious way to do this would be to install tracking software on the participants' smartphones, recording the data, and allowing the researchers to label the data themselves. Besides being invasive, this would likely alter user behavior due to the Hawthorne effect. An additional problem is the burden of labeling for numerous participants. It is likely not feasible for a small team of researchers to label thousands to millions of data points themselves. The temptation would be to outsource the labeling. For instance, crowdsourcing could be used to label visual data collected in-the-wild. In many cases, this will violate privacy rights and again, participants would likely alter their daily habits if they are notified of this labeling of personal data during their informed consent. Moreover, behavioral data may need to be labeled by annotators who have sufficient domain expertise to accurately label the data,[8] or rely on a process that includes enough data so that mistakes by annotators can be acceptable.[9]

The other option is to have participants label data by themselves. This has the advantage of allowing

participants to act as a filter over their own data submission, thus maintaining privacy and dignity. However, asking participants to label data could be quite burdensome and alter their behavior. Interleaved labeling, where the participant has to interrupt their day to provide labels, is difficult, and could also introduce a modifying effect on their behavior. In some cases, especially those seeking for experience sampling, researchers provide interruptions via smartphone notifications, which are not always answered.[10] This directly interferes with the participants' natural behavior. Other systems may leave it up to the participant to label data as they go or after specific actions. This may cause interruptions in natural behavior. Participants may avoid a certain behavior because they feel they have labeled that behavior previously. The better option is to have participants provide data labels after a specific collection period. In order to make this possible, there needs to be a mechanism to facilitate recall.

PRELUDE STUDY: DESIGN

All of the experiments and work described in this article were done with the permission of the research ethics committee of the Graduate School of Engineering, Osaka Prefecture University.

In order to present our method for data labeling in-the-wild, we first introduce the background study that motivated its development. In this study, we aimed to differentiate between *controlled* reading and *natural* reading.[11] In order to do this, we recruited seven participants to record their habits using J!NS MEME glasses, off-the-shelf consumer glasses equipped with a dry-electrode setup that allows for the recording of eye movements through electrooculography, and Narrative Clip, which captures still pictures every 30 seconds (the device can be attached to the front of clothing and provides a point-of-view recording). In total, dataset contains 22 hours of controlled reading, 427 hours of natural reading, 156 hours of social interactions, and 375 hours of other activities. This study was modestly successful in implementing participant self-labeling.

Natural reading activity: No limits were placed on the participants' reading activities. Anything from reading on computers, smartphones, or paper was classified as natural reading. In order to make the classification more discrete, reading was defined as purposefully reading blocks of text. Reading of road signs or text in movies did not count as natural reading.

Participants added annotations to all data in the natural reading activity by using the pictures collected from a wearable camera worn on the front of their shirts called the Narrative Clip. Participants applied one of the three labels ("reading," "talking," and "other activities") to every 1 minute of data from 0:00 to 23:59. To reduce ambiguities, we asked participants to label activities if pertinent objects (e.g., book, display, person) appeared in more than two consecutive pictures (= one minute). Labeling activities were asked to be completed at the end of each day after recording to minimize the interruptions of participants' everyday lives, and to ensure that recall of the activities would be possible. Participants were encouraged to remove pictures that they did not want to share with the researchers in order to protect their privacy.

Controlled reading activity: We also conducted a controlled experiment where we prepared 60 documents, split between computer and paper documents. Participants were asked to read them from beginning to end during the course of the day. This allowed us to have a base-level to compare the natural reading activity dataset. Reading on paper was recorded with J!NS MEME, and reading on a screen was recorded with J!NS MEME and Tobii eyeX.

Prelude Study: Findings

The results of the classification algorithms showed that there were some problems with false positives and false negatives in the natural reading activity analysis. By reviewing the pictures taken by Narrative Clip, we identified some cases in which an activity was misclassified by the participants themselves. This was primarily due to two reasons: 1) The act of reading differs (e.g., reading a paper book, browsing web pages, skimming texts, etc.), and classifying activities into two simple classes (reading versus not reading) can sometimes be difficult even for humans, and 2) labeling the data is very time intensive for the participants due to the sheer amount of pictures without a convenient workflow.

Another problem was highlighted by the study. Participants were allowed to curate their pictures from the Narrative Clip, but this puts pressure on them to make a decision whether or not to redact a picture for privacy. This is problematic as it creates an opportunity for participants to make mistakes in the data curating stage; mistakes that are more likely to occur when the data points and the amount of labels needed are high. In larger scale studies, the chances of inadvertently harming a participant's dignity increases greatly (e.g., the participant inadvertently does not redact an image due to the number of images they are required to curate). Thus, the burden on participants needs to be lowered.

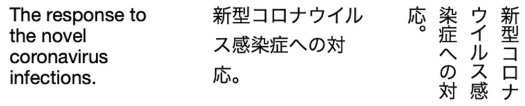

FIGURE 1. Examples of different reading layouts: English, Japanese Horizontal, and Japanese Vertical. All three texts have identical meanings.

DATA LABELING IN-THE-WILD

Using the lessons from the aforementioned study, we developed a process and tools to help participants self-labeling their data. First, we developed the Narrative Logger,[a] which seeks to aid the participants in quickly labeling their data. We also provided the participants with directions facilitating the labeling process.

Goal of the Study

There is diversity in how documents are presented and there are numerous reading styles in the world. For instance, a Japanese university student would not only deal with reading on different mediums (e.g., computer, smartphones, books, etc.), but would also be reading English, Japanese presented in horizontal layout going left-to-right, and Japanese presented in vertical layout going right-to-left. Figure 1 shows an example of a sentence written in the three layouts (the English is a translation of the Japanese, while the two Japanese layouts have identical characters). Because of this diversity, developing inclusive reading detection algorithms is difficult.

To address this problem, we developed an experiment where the participants would read in all three styles in an in-the-wild setting.[12] To do this, we recruited 10 Japanese university students to record their reading habits for two days. They were requested to record for 10 hours each day, but no strict limitation was imposed. The devices used in this study were the J!NS MEME glasses, which records electrooculography, Android Nexus 5X smartphone with custom logging software for participant control,[b] and the Narrative Clip 2, which recorded pictures of what the participant was doing.

We explicitly asked the participants to read three types of text formats [English (EN), Japanese Written Horizontal (JH), and Japanese Written Vertically (JV)] for one hour each for each day. There was no restriction placed on how these three hours should be accomplished and no time-recording requirement for the participants. At the end of each day, the participants labeled their data with the pictures recorded from the Narrative Clip 2. The participants provided one of four labels: EN, JH, JV, or Not Reading (NR). Unlike the previous study, the participants only provided labels, instead of pictures, to safeguard their privacy. We provided the participants with a custom software tool to help them quickly label their pictures. The participants received 10,000 JPY (approx. 100 USD) for their participation and the completion of their tasks.

Development of Narrative Logger

The Narrative Clip 2 took photos every 30 seconds during the hours of recording. At the end of the day, if participants followed the 10 hours request, they would have 1200 photos to label manually. As shown in Figure 2, there were four categories that the participants could choose. As mentioned before, there was a design choice to be made in the labeling workflow; we could either have the participants provide labels interleaved with their activities or require labeling at the end of the day.

We chose the latter to prevent the continuous disruption of our participants' everyday lives and encourage natural behavior. The design discussion around the interleaved labeling allows for two options: 1) the participants are given directions to provide labels immediately after readings, and 2) the participants are allowed to label whenever they want (the labeling may be interleaved with activities, but also may not). The first option has some positive attributes. Participants are less likely to incorrectly label data points if they are labeling at the same time. The main drawback is the disruption of natural behavior. For instance, a participant may refrain from a reading activity because of the added time and effort of labeling. This would also encourage task binning; a participant does all of one task at once to get the labeling out of the way. The second option, while allowing more freedom, may cause disruption during the day, creating an imbalance in when and how data points are labeled. Participants would have to remember specific time periods that they have missed in their labeling tasks if they decide to label their pictures occasionally interleaved with their activities.

The option to label pictures at the end of the day accommodates different levels of busyness that a participant may face. While it might be possible to label periods of reading when one is having a relaxed day, a hectic day may make it impossible to be exact. Having

[a]https://narrativelogger.shoya.io
[b]https://memelogger.shoya.io/

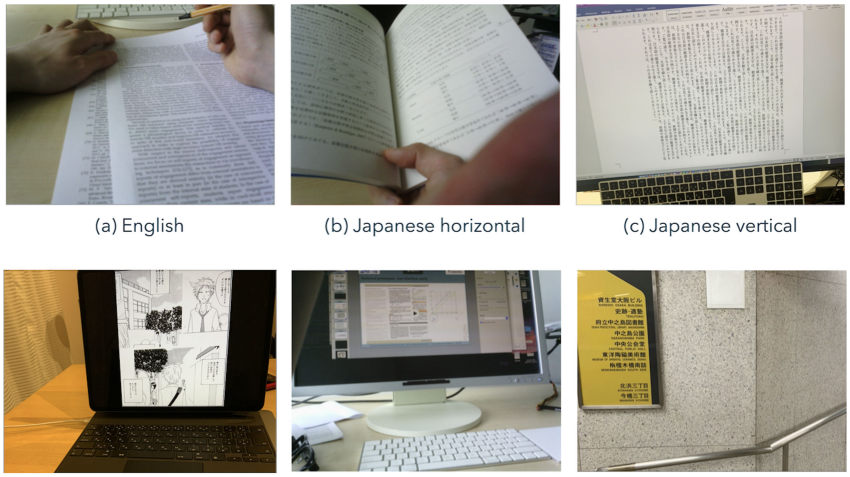

FIGURE 2. Examples of reading and not reading.

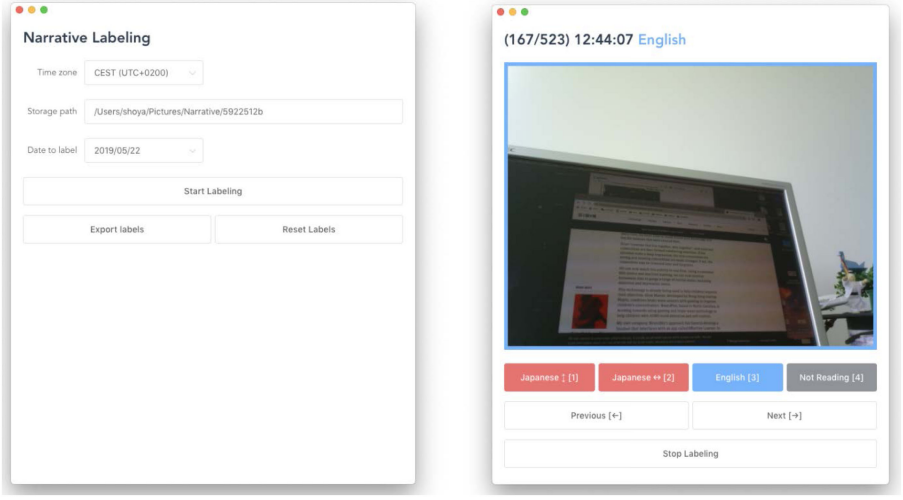

FIGURE 3. Narrative Labeling tool.

pictures from the Narrative Clip 2 to review days provides tangible references for the participant to correctly label their data. The potential drawbacks to this option is that labeling at the end of the day is burdensome and potentially time-consuming.

Based on this logic, we developed the Narrative Labeling tool focusing on a linear progression of pictures that were to be labeled at the end of the day. As shown in Figure 3, the labeling tool provides the pictures taken by the Narrative Clip 2 in chronological order. This allows the participants to quickly recognize blocks of time where activities and related activities take place. The participants could then scroll through the pictures using arrow keys and dedicated hot keys to label the pictures.

Participants were instructed to label the data using the four labels (EN, JH, JV, and NR) with strict definitions. A picture needed labeling not only if there was text in the captured photo, but also only if the participant was engaged in reading at that time. However, we instructed users to only label text that requires line-breaks. As shown in Figure 2, even if text characters appeared in a picture, if the text does not require frequent line breaks, it was not categorized as reading in this study. Therefore, looking at a street-sign, reading a comic, or writing code was not classified as reading in this particular study.

A key aspect of the development of the program was the amount of control over the picture data given to the

FIGURE 4. Time spent by each user on tasks for both days. Colors represent the relationship to the requested time spent on each task and percentage of deviation from request.

participants. Participants only needed to submit a CSV file of their labels, and were informed that the pictures from the Narrative Clip 2 would be deleted from their device as soon as the labels were recorded. Therefore, we only interacted with the labels, not the pictures themselves.

Participants received an initial briefing on the labeling process, written instructions, and could contact the experimenters directly for help. Beyond this, direct interaction with the participants was limited to reduce the observer effect.

Results and Takeaways

We extracted the following 10 features that were calculated from one sample (30 seconds window of a data stream): means and variances of the two EOG axes, variances of three accelerometer axes, and variances of three gyroscope axes. Then, we utilized a support vector machine (SVM) to classify the samples. The radial basis function kernel with hyperparameters C = 1 and $\mathrm{gamma} = 0.125$ were selected for the classifier based on an iterative approach.

We used both user-independent and a user-dependent training. We used leave-one-participant-out cross-validation for user-independent training. For the user-dependent training, the classifier was adapted to each participant with their data with leave one-day-out cross-validation. Since the dataset was unbalanced, we applied undersampling for the evaluation.

We achieved promising results. SVM was able to provide much better-than-chance classification (which was set to 50%) for each type of reading versus not reading in both user-independent and user-dependent classifications (the smallest difference was 16% improvement and the largest 24%). For this article, however, the most important aspects of the study were in how participants used and reacted to the Narrative Logger.

As Figure 4 shows, participants typically logged and labeled more data than required. In total, the dataset contains 23 hours of English reading, 25 hours of Japanese horizontal reading, 25 hours of Japanese vertical reading, and 146 hours of other nonreading activities. We received more data for three reading classification activities in 37 out of 60 possible daily cases, and only received less than 10% of the required 60 minutes in three cases. This is a surprising and promising result, as it shows that participants went beyond what they were asked and remunerated to do.

The Narrative Logger is responsible for this result. While exact numbers were not recorded, participants reported that labeling an entire day of data took only 10–15 minutes. The layout of our Narrative Logger facilitated quick and easy labeling. A key element here is that number of reading events throughout the day is much lower than the number of pictures. Labeling at the end of the day allowed participants to quickly recall what they were doing in the pictures.

As one participant noted, the labeling was easily done at the end of the day with the tool provided; he could label data outside of the three required hours and focus on reading.

DISCUSSION

The development of the Narrative Logger and the accompanying directions provide a proof-of-concept for data labeling in-the-wild. Participants can and will label their own data if provided with the correct tools. Although recall may be considered to be error-prone for labeling accuracy and pictures too time-consuming,[13] the combination of the two is effective.

Our goal in developing this methodology is to provide a way of obtaining accurate data labels without setting up artificial environments, interrupting the natural behavior of study participants, or exposing participants to situations in which their privacy can be violated. By collecting these data outside of the lab, researchers are able to reduce the amount of direct human contact needed to run their studies, not only following recent restrictions to social interaction, but also enabling remote studies beyond local communities.

Related Approaches

Our methodology is not the first attempt to facilitate data labeling in-the-wild. Vaizman[14] proposed *ExtraSensory*, a mobile application for collecting data labels. One of the key contributions of ExtraSensory is providing an understanding of how labels can be given for a wide range of activities with different labeling options. In particular, participants are able to label data interleaved with their activities, e.g., via notifications, or after their activities (history). It turns out that history is much more popular among the participants, possibly due to the interruption that interleaved labeling causes. The authors propose that cameras, like the ones we employed, could be used to aid this historical labeling.

Alharbi[15] used a camera to record video for an experiment where users labeled whether or not they were eating after participating in three to four hours of activities. The authors developed a scanning tool to expedite the labeling process. This method may be appropriate for situations where the camera is capturing the participant in the frame rather than being deployed as a point-of-view device. For instance, the users in the eating study may not need to see a point-of-view picture to remember when they were or were not eating. A limitation to the video element is the amount of time it takes to label the video. In this study, the amount of recorded footage was limited to a few hours at a time.

Devices Used

The two devices that are employed in both studies also merit discussion. We asked participants, who do not normally wear glasses, to wear the J!NS MEME devices. While the J!NS MEME has been successfully used in long-term in-the-wild studies,[16] there is a consideration that attention should not be brought to the device by the researchers if possible. That is, participants begin to wear the device naturally as time goes forward. Therefore, interleaved labeling would likely bring more attention to the wearable. As the electrooculography glasses are not related to the participant tagging activity, it makes sense to have this device removed from the process.

The Narrative Clip 2 proposes an issue for third-party nonparticipants and interaction with the participants. Social acceptability of device usage needs to be considered carefully.[17] The concept of recording devices entering public spaces has long been an area of concern in pervasive computing research.[18] If participants of a study are asked to wear a front-facing camera like the Narrative Clip 2, this may induce a bystander effect where third parties who recognize the device react negatively toward the participant. While there may be ways to decrease the bystander effect by obscuring the camera,[19] the reduced view could harm labeling quality in many cases. However, the context of a particular study must be considered when deciding how to deploy the wearable devices. Using cameras or videos for wide-ranging behavioral studies that seek to capture many different behaviors may introduce behavioral changes and violate the privacy of nonparticipants.[20]

Implications

The implications of the Narrative Logger and its study are twofold. First, the tool set is can be readily used for projects that require visual data labeling in-the-wild. While the tool-setup may need to be fine-tuned for a particular study, a workflow is available to anyone who downloads the software and acquires the Narrative Clip 2 devices. In addition, the Narrative Logger is published as an open-source software. Allowing researchers to easily modify it to work with other camera systems.

The larger scale implication is that we achieved data labeling by the participants themselves at the end of the day. What seems like an onerous task for participants becomes one that can be done quickly by using pictures as an anchor-point. This could be employed in situations where activities that are conducted in-the-wild need to be labeled and disruptions to the participants' daily lives are not desirable. This opens up the possibility for similar workflows for other types of labeling beyond activities. For instance, pictures could be used to help study participants recall emotions throughout the day.

This development is important as pervasive computing becomes more and more important in everyday life. The proposed workflow not only reduces the need for in-person researcher–participant contact, which is difficult under COVID-19 restrictions, but it also provides the participant with more control over their data-stream.

Limitations

A limitation to our work is that the positive results obtained by the Narrative Logger are largely due to the careful construction of the program within a well-defined project space and context. Because the authors have been working in and analyzing their domain of study—reading detection—there is a strong sense of what will and will not work. In addition, the workflow may be strongly suited to this context. In other contexts and situations, the workflow and synergy between technologies may not be so clear and may need fine-tuning.

Another limitation is the medium required to record the labels. At present, the proposed workflow only works with visual labeling. The ability to use linear progression and keyboard shortcuts allows participants to label their data quickly and easily due to the speed of visual recognition. This would be difficult for audio or extended video labeling, where participants have to spend time on playback.

The labeling technique that was presented in this article was not directly compared against other types of labeling (such as interleaved labeling). There is a possibility that the workflow presented here may be augmented to be more effective. This remains part of future work.

Another consideration is the amount of labels that are employed in the study. Our workflow depends on four discrete labels. In some research, there may be a need to provide more granular or overlapping labels. This would likely increase the error rate and the amount of time needed for providing labels and training the participants. A necessary step for future work is to explore the limits of what task instructions are feasible under the workflow.

Finally, further research needs to be conducted on the amount of the mitigation of the Hawthorne effect. While we expect that the observer influence is mitigated due to participant control, there may be a pseudo-Hawthorne effect that is still present. This is an important question going forward for the future of all in-the-wild studies.

CONCLUSION

In this article, we provide a method for obtaining data labels from studies that are run in-the-wild. As the need for trustworthy labeled data that reflect natural behavior increases, it is necessary to innovate ways in which to collect these labels from participants. We propose a workflow that uses an application that allows users to independently and privately log their data labels after a data recording period.

As the world slowly recovers from COVID-19, there is a chance to permanently run more studies out of the lab. If data labels can be gathered without disrupting the daily lives of participants, moving to in-the-wild situations could prove to be beneficial for the technologies being developed by reflecting more natural behaviors. With a workflow that promotes user control and relies on participant recall, it is feasible to collect these labels.

ACKNOWLEDGMENTS

This work was supported in part by grants from JST CREST under Grant JPMJCR16E1, JSPS Grant-in-Aid for Scientific Research (B) under Grant 20H04213, and Grand Challenge of Initiative for Life Design Innovation, MEXT Society 5.0 Realization Research Support Project, Osaka University.

REFERENCES

1. A. Chamberlain, A. Crabtree, T. Rodden, M. Jones, and Y. Rogers, "Research in the wild: Understanding 'in the wild' approaches to design and development," in *Proc. Designing Interactive Syst. Conf.*, 2012, pp. 795–796, doi: 10.1145/2317956.2318078.
2. B. Brown, S. Reeves, and S. Sherwood, "Into the wild: Challenges and opportunities for field trial methods," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2011, pp. 1657–1666, doi: 10.1145/1978942.1979185.
3. O. D. Lara and M. A. Labrador, "ASurvey on human activity recognition using wearable sensors," *IEEE Commun. Surv. Tut.*, vol. 15, no. 3, pp. 1192–1209, Jul.–Sep. 2013, doi: 10.1109/SURV.2012.110112.00192.
4. A. E. Cunningham and K. E. Stanovich, "What reading does for the mind," *Amer. Educator*, vol. 22, pp. 8–17, 1998.
5. F. Larradet, R. Niewiadomski, G. Barresi, D. G. Caldwell, and L. S. Mattos, "Toward emotion recognition from physiological signals in the wild: Approaching the methodological issues in real-life data collection," *Front. Psychol.*, vol. 11, 2020, Art. no. 1111, doi: 10.3389/fpsyg.2020.01111.
6. J. McCambridge, J. Witton, and D. R. Elbourne, "Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects," *J. Clin. Epidemiol.*, vol. 67, no. 3, pp. 267–277, 2014, doi: 10.1016/j.jclinepi.2013.08.015.
7. S. Reyal, S. Zhai, and P. O. Kristensson, "Performance and user experience of touchscreen and gesture keyboards in a lab setting and in the wild," in *Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst.*, 2015, pp. 679–688, doi: 10.1145/2702123.2702597.
8. A. Joshi, S. Kyal, S. Banerjee, and T. Mishra, "In-the-wild drowsiness detection from facial expressions," in *Proc. IEEE Intell. Veh. Symp.*, 2020, pp. 207–212, doi: 10.1109/IV47402.2020.9304579.

9. H. M. S. Hossain, M. A. A. H. Khan, and N. Roy, "Active learning enabled activity recognition," *Pervasive Mobile Comput.*, vol. 38, pp. 312–330, 2017, doi: 10.1016/j.pmcj.2016.08.017.
10. N. van Berkel, J. Goncalves, S. Hosio, Z. Sarsenbayeva, E. Velloso, and V. Kostakos, "Overcoming compliance bias in self-report studies: A cross-study analysis," *Int. J. Hum.-Comput. Stud.*, vol. 134, pp. 1–12, 2020, doi: 10.1016/j.ijhcs.2019.10.003.
11. S. Ishimaru, K. Hoshika, K. Kunze, K. Kise, and A. Dengel, "Towards reading trackers in the wild: Detecting reading activities by EOG glasses and deep neural networks," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, 2017, pp. 704–711, doi: 10.1145/3123024.3129271.
12. S. Ishimaru, T. Maruichi, M. Landsmann, K. Kise, and A. Dengel, "Electrooculography dataset for reading detection in the wild," in *Adjunct Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, 2019, pp. 85–88, doi: 10.1145/3341162.3343812.
13. M. Stikic, D. Larlus, S. Ebert, and B. Schiele, "Weakly supervised recognition of daily life activities with wearable sensors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2521–2537, Dec. 2011, doi: 10.1109/TPAMI.2011.36.
14. Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel, "ExtraSensory App: Data collection in-the-wild with rich user interface to self-report behavior," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2018, pp. 1–12, doi: 10.1145/3173574.3174128.
15. R. Alharbi, T. Stump, N. Vafaie, A. Pfammatter, B. Spring, and N. Alshurafa, "I can't be myself: Effects of wearable cameras on the capture of authentic behavior in the wild," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–40, 2018, doi: 10.1145/3264900.
16. B. Tag, A. W. Vargo, A. Gupta, G. Chernyshov, K. Kunze, and T. Dingler, "Continuous alertness assessments: Using EOG glasses to unobtrusively monitor fatigue levels in-the-wild," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2019, pp. 1–12, doi: 10.1145/3290605.3300694.
17. M. Koelle, S. Ananthanarayan, and S. Boll, "Social acceptability in HCI: A survey of methods, measures, and design strategies," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2020, pp. 1–19, doi: 10.1145/3313831.3376162.
18. S. Neely, G. Stevenson, C. Kray, I. Mulder, K. Connelly, and K. A. Siek, "Evaluating pervasive and ubiquitous systems," *IEEE Pervasive Comput.*, vol. 7, no. 3, pp. 85–88 Jul.–Sep. 2008, doi: 10.1109/MPRV.2008.47.
19. R. Alharbi, M. Tolba, L. C. Petito, J. Hester, and N. Alshurafa, "To mask or not to mask? Balancing privacy with visual confirmation utility in activity-oriented wearable cameras," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–29 2019, doi: 10.1145/3351230.
20. Y. Vaizman, K. Ellis, and G. Lanckriet, "Recognizing detailed human context in the wild from smartphones and smartwatches," *IEEE Pervasive Comput.*, vol. 16, no. 4, pp. 62–74, Oct.–Dec. 2017, doi: 10.1109/MPRV.2017.3971131.

ANDREW VARGO is currently a Project Assistant Professor at Osaka Prefecture University, Sakai, Japan. His research interests include the creation of collaborative systems for learning, health, and well-being. He is the corresponding author of this article. Contact him at awv@m.cs.osakafu-u.ac.jp.

SHOYA ISHIMARU is currently a Junior Professor with the Department of Computer Science, University of Kaiserslautern, Kaiserslautern, Germany, and the chief research officer of Alphaben. His research focuses on inventing technologies augmenting human intellect. Contact him at ishimaru@cs.uni-kl.de.

MD. RABIUL ISLAM is currently a doctoral student at Osaka Prefecture University, Sakai, Japan, and an assistant professor at BSMRSTU, Gopalganj, Bangladesh. His research focuses on deep learning technologies for recognizing human cognitive activities. Contact him at rabiul@m.cs.osakafu-u.ac.jp.

BENJAMIN TAG is currently a Postdoctoral Researcher at the School of Computing and Information Systems, University of Melbourne, Parkville, VIC, Australia. His research interest is human cognition, with focus on inferring cognitive state changes from biophysical signals. Contact him at benjamin.tag@unimelb.edu.au.

KOICHI KISE is currently a Professor with the Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai, Japan. His research interests include human activity recognition, human–computer interaction, document analysis and interaction, and image analysis and recognition. He has been an IEEE member for 20 years, IEEE Computer Society member for 31 years, and IEEE Signal Processing Society member for 16 years. Contact him at kise@cs.osakafu-u.ac.jp.