SenseCollect: We Need Efficient Ways to Collect On-body Sensor-based Human Activity Data!

WENQIANG CHEN, University of Virginia
SHUPEI LIN, VibInt AI Limited
ELIZABETH THOMPSON, University of Virginia
JOHN STANKOVIC, University of Virginia

On-body sensor-based human activity recognition (HAR) lags behind other fields because it lacks large-scale, labeled datasets; this shortfall impedes progress in developing robust and generalized predictive models. To facilitate researchers in collecting more extensive datasets quickly and efficiently we developed SenseCollect. We did a survey and interviewed student researchers in this area to identify what barriers are making it difficult to collect on-body sensor-based HAR data from human subjects. Every interviewee identified data collection as the hardest part of their research, stating it was laborious, consuming and error-prone. To improve HAR data resources we need to address that barrier, but we need a better understanding of the complicating factors to overcome it. To that end we conducted a series of control variable experiments that tested several protocols to ascertain their impact on data collection. SenseCollect studied 240+ human subjects in total and presented the findings to develop a data collection guideline. We also implemented a system to collect data, created the two largest on-body sensor-based human activity datasets, and made them publicly available.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *on-body sensors*; human activity recognition; • **Data collection** → micro finger writing.

Additional Key Words and Phrases: Human activity recognition, Data collection, On-body sensors, Micro finger writing

ACM Reference Format:
Wenqiang Chen, Shupei Lin, Elizabeth Thompson, and John Stankovic. 2021. SenseCollect: We Need Efficient Ways to Collect On-body Sensor-based Human Activity Data!. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 1 (July 2021), 27 pages. https://doi.org/10.1145/xxx.xxxxxx

1 INTRODUCTION

On-body sensor-based human activity recognition (HAR) is a vital research area that is lagging behind its compatriots. HAR is widely used for a variety of applications, from human computer interaction and user authentication-based security systems to healthcare uses such as Parkinson disease prediction and fitness tracking wristbands [5, 7, 32, 42, 55] as shown in figure 1. However, despite its utility the HAR field has yet to experience significant improvements in activity recognition performance; this is in stark contrast to breakthroughs evident in other fields [29], such as speech recognition [25], natural language processing [17], and computer vision [23]. A key difference that separates HAR from those fields is that each of those domains has significantly large amount of labeled data. For instance, the ImageNet dataset [16] has approximately 14 million images, and the "One

Authors' addresses: Wenqiang Chen University of Virginia, wc5qd@virginia.edu; Shupei Lin VibInt AI Limited; Elizabeth Thompson University of Virginia; John Stankovic University of Virginia.

2474-9567/2021/7-ART $15.00
https://doi.org/10.1145/xxx.xxxxxx

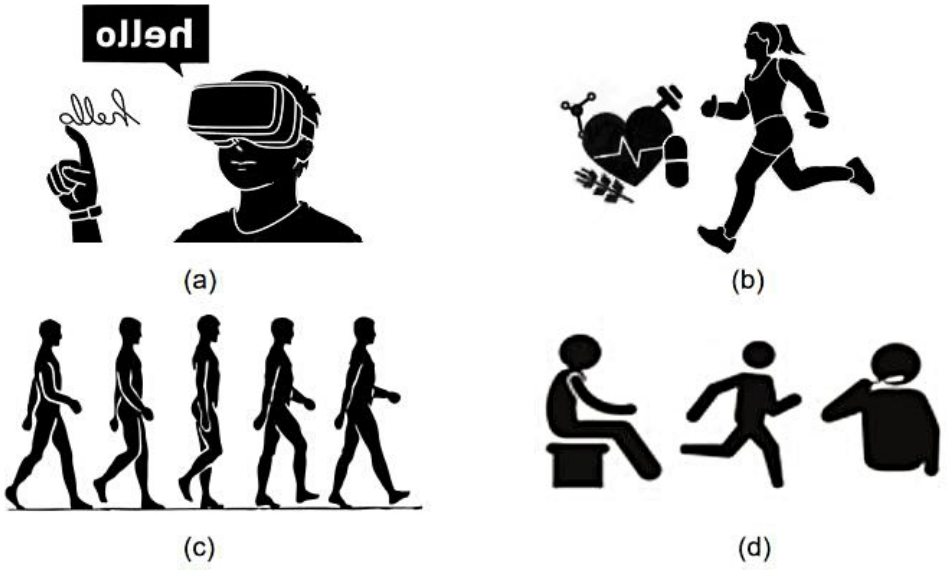

Fig. 1. A wide variety of HAR applications: (a) Human-computer interaction (b) healthcare (c) gait authentication (d) tracking everyday activities

billion words" benchmark [8] contains literally one billion words, and both datasets are publicly available. Large datasets are vital for deriving robust recognition models that strongly generalize across application boundaries. In contrast HAR does not have the same magnitude of data; generally the scale of human activity datasets are small, thereby only covering limited sets of activities with insufficient data for generalization [7, 26, 46, 55]. Even the largest, accessible, sensor-based activity dataset only spans a few users with relatively short durations [5, 40]. For example, the Daphnet freezing of gait dataset [5] only has 5 hours of sensor data in total, collected across 10 participants, and the PAMAP2 dataset [40] has 7.5 hours of sensor data collected across 9 participants. Models derived from sparse datasets are less likely to generalize well, especially when compared with models created using larger datasets. Deep learning speech transcription models are a common household item these days, they exist on most smartphones and smart home devices, and are usable by anyone, including voices transmitted from speakers or a television. Opposite that, most HAR models have inferior performance when generalized to other users. Researchers are hampered by the lack of available data, making it difficult to create models that perform well across other users. Furthermore, only a limited amount of activity datasets, usually coarse-grained, are made publicly available. These kinds of datasets are not adequate for conducting new and fine-grained HAR research, such as mid-air micro finger writing recognition.

In order to achieve the breakthroughs experienced by other fields, like computer vision and speech recognition, HAR needs greater availability of data. Therefore, it is imperative that researchers working with on-body sensor-based HAR collect more datasets. However, on-body sensor-based HAR data is challenging to collect and label, especially compared to data in other domains, like computer vision, where data is easily accessible. Since HAR data is not readily available creating new datasets often requires researchers to design and implement human research studies, making that data laborious, complicated, and error-prone. (see section 3.2.1) During our surveys and interviews in the preliminary study (Section 3) all student researchers cited data collection as the greatest barrier in their research. One fourth-year Ph.D. candidate (P4) lamented that he spent two years solely on

collecting data only to have the majority of the datasets rendered useless due to high levels of error. Many students stated that they had difficulty finding participants and often had to beg their friends to participate in their data collection experiments. This kind of situation puts additional stress on the researchers, especially when a friend might get frustrated and quit the experiment partway through, leaving the researcher with no data and an annoyed friend. Additionally students observed that impatient participants generally had lower accuracy data, and that all participants frequently made unconscious motions which polluted the dataset. Recently collecting data has become incredibly difficult due to the COVID-19 pandemic, and many interviewees' research projects have been significantly delayed due to their inability to meet participants for data collection in many countries. One interviewee (P10) mentioned that he has given up on research in the area of on-body sensor-based HAR in favor of natural language processing (NLP), citing the exhausting data collecting experiments required for HAR research as the primary reason he switched fields.

We have identified three main challenging aspects of on-body sensor-based data collection:

1 Collecting activity data with sensors from human bodies is tedious, laborious and time-consuming.
2 It is difficult to collect activity data that is subsequently used for manual data annotation since the on-body sensor's signals are not easily understood by people.
3 Since on-body sensor-based human activities do not have sensing feedback, such as vision or audio, the datasets are easily polluted by deformed activities and unconscious motions, thus being error-prone.

Given the above challenges, how can we facilitate efficient data collection for HAR researchers? The HAR field needs more datasets, but how can we assure the data quality and prevent data pollution from participants' different personalities and unconscious motions?

In this paper we present SenseCollect, a framework that includes data collection guidelines and software to collect on-body sensor-based HAR data. SenseCollect was developed based on the results from a three-phase study designed to address the above questions and explore ways to overcome the challenges of data collection. In order to improve the quality and quantity of HAR data available it is necessary to attain a better understanding of what the influential factors in HAR data collection are, with the goal of swiftly increasing data collection so that a large on-body sensor-based HAR dataset collection can be achieved. First, we did a preliminary study where we did a survey and interviewed student researchers in HAR research who have had experience with on-body sensor-based human activity data collection. We then used that data to explore efficient data collection by developing a system to collect data and conducting a series of control variable experiments to investigate the impact that different ways of organizing participants and different potential factors can facilitate or complicate the data collection. In the third phase, we conducted a study that allowed us to simultaneously explore the impact that several factors have on data collection while also creating two new HAR datasets which are, to the best of our knowledge, the two largest HAR datasets available.

Our studies involved over 240 participants in total, allowing us to gather information and data from a variety of different people. Over the course of the study we discovered several different factors that potentially impact the data collection process, both the data quality and the amount collected. Later in our paper we present our findings and discuss how their influence can be utilized to improve research techniques going forward. As a culmination of our experiments, SenseCollect presents guidelines on how different ways of organizing human participant studies can affect data collection, as well as what factors may impact collecting data. Factors tested include using sensing feedback, providing techniques to address the effect of the psychologically uninteresting nature of the data collection, imposing data quality requirements, using people monitoring, and setting time limits on data collection periods.

By combining social science techniques, control variable experiments, and assistance-based system development, SenseCollect broke through the various obstacles of on-body sensor-based HAR dataset collection and found an efficient way to collect large and accurate datasets. SenseCollect is suitable for collecting data during a

pandemic. 5.1 Also, we collected two on-body sensor-based HAR datesets successfully, based on our findings. These datasets are publicly available and it can be downloaded at https://www.jianguoyun.com/p/DVaDq-QQ2enUCBiLx-AD. We believe that they are the largest on-body sensor-based HAR datasets available, and include 192 hours of finger writing data from 32 subjects, and 113 participants' data of on-body tapping. Note that SenseCollect is meant to get ground truth training data and so this data must be correct (and labeled as such). We trained models for these two datasets and obtained average validation accuracy at about 90% and 91% respectively. We also evaluated the training dataset quality with real-time test data in practice with a user study (see section 6.4). We believe these datasets are able to contribute to many future research projects. (see section 6.5). Additionally we developed a system for Windows, Mac and Android computers to assist in the data collection; this software can be downloaded at the same link. Note that we do not claim SenseCollect is the best way to collect on-body sensor-based HAR datasets, however it is a significant improvement over today's methods.

In summary, the major contributions of this paper are as follows:

- SenseCollect is able to obtain large, accurate datasets collected in a short time period of days or weeks, where as current practices often take months and frequently result in small datasets that are error-prone.
- We investigated, identified, and experimented with the difficulties for on-body sensor-based human activity data collection with 241 subjects and found an efficient way to collect much larger and accurate datasets than currently exists.
- Using information gathered by a preliminary study, we identified and analyzed how key factors mentioned above influence HAR data collection. To study these factors we used a highly demanding HAR data collection, i.e., finger writing in the air based on vibrations through the hand to a smart watch.
- After analyzing the effects of these factors, We created a data collection assistance system and collected the two largest on-body sensor-based HAR datasets that exist from 163 participants. The data collection system and the datasets are publicly available.
- Additional experiments were conducted with these new datasets using neural networks, such as GRU-CTC models, resulting in about 90% accuracy.

2 STUDY PROCEDURE

In order to facilitate student researchers in collecting larger on-body sensor-based HAR datasets efficiently and easily, SenseCollect conducted a study consisting of three phases as shown in Figure 2. To understand the difficulties of collecting on-body sensor-based human activity data, we began our first phase study which used surveys and semi-structured interviews from student researchers. We were interested in recruiting students working in on-body sensor-based human activity recognition and who have had experience collecting data for their research. From this study we determined several potential factors affecting data collection. Based on this preliminary study, in phase two we designed further experiments to investigated what and how the affecting factors may complicate the data collection. Here we used a quasi-experimental design to test which factors potentially influence data collection. In phase three we implemented the experimental design using the identified factors as independent variables, and then we analyzed our results and developed guidelines for data collection based on our findings. The studies involved 241 human subjects in total, including three researchers, 50 survey participants, five in group A, eight in Group B, 62 in Group C, and 113 in generalized experiments. Note that all of the experiments involving human subjects conformed to the relevant regulations of our institute. The details for each phase are described below.

- Phase 1- Preliminary study: We did a survey from 50 student researchers and interviewed ten of them to understand the difficulties of on-body sensor-based human activity data collection. From this interview, we identified difficulties in conducting human involved experiments. Additional challenges identified included

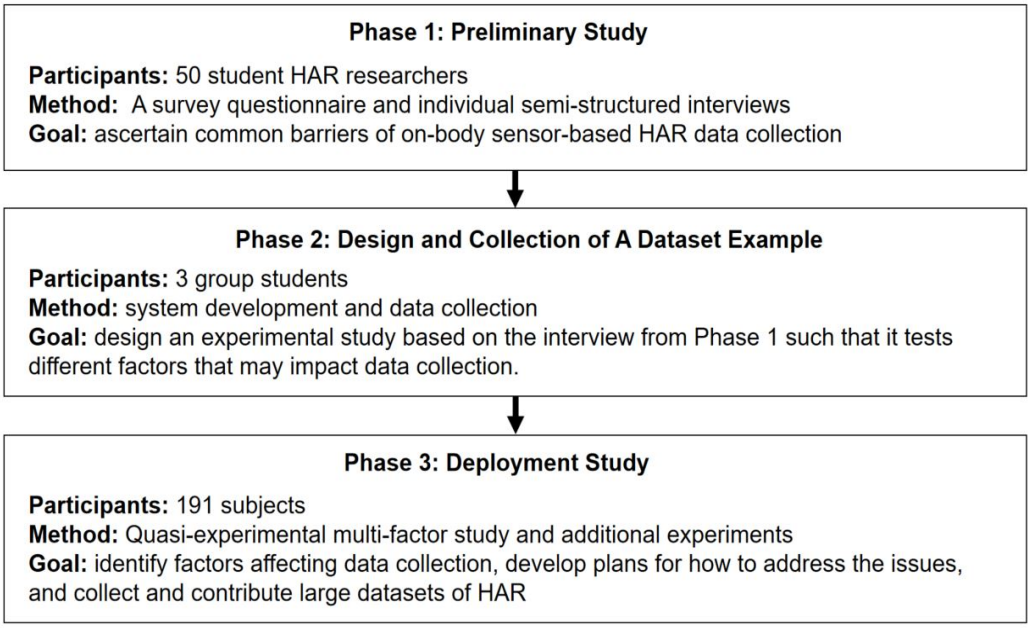

Fig. 2. Overall study procedure

that on-body sensor-based datasets are error-prone, and that data collection is tough for student researchers. This preliminary study showed why accessible sensor-based HAR datasets are small.

- Phase 2 -Design and collection of a dataset example: We designed an on-body sensor-based human activity dataset and conducted experiments to collect it. In this way, we achieved the biggest on-body sensor-based human activity datasets so far. We first investigated three different ways to organize participants to collect their activities. We recruited two student researchers working in this area (group A), eight lab mates not working in this area (group B), and 18 random undergraduate investigators (group C). In group C, each investigator was required to collect three other participants' HAR data. Then, for Group C, we further explored what factors resulted in the difficulty of collecting on-body sensor-based human activities by a series of control variable experiments. We present the design details in section 4.
- Phase 3 - Deployment study: This section presents our findings on the impact of various factors, including: using different ways to organize participants, using sensing feedback, addressing psychologically uninteresting nature, the effect of a data quality requirement, the value of monitoring, and the length of data collection period. A better understanding of these factors is central to achieve larger on-body sensor-based human activities data collection. Based on the control variable experiments, we present how we overcame difficulties and found an efficient way to collect data. We were able to achieve the biggest on-body sensor-based human activity datasets, which we have made publicly available for further research.

3 PRELIMINARY STUDY

In our preliminary study, we investigated the current practice of on-body sensor-based HAR data collection by a medium-large scale survey and interviewing some of the survey participants.

Table 1. Student researchers and their research applications.

Interviewee	Program	Applications
P1	Ph.D.	Mid-air micro finger writing recognition
P2	Ph.D.	Finger gestures recognition,
P3	Ph.D.	Hand-to-hand gestures recognition
P4	Ph.D.	Sign language translation
P5	Ph.D.	In-air hand writing for authentication
P6	Ph.D.	Fitness tracking
P7	Ph.D.	HAR accuracy with different sensors
P8	Master	Hand gestures recognition for military
P9	Master	Sleep monitoring
P10	Undergraduate	HAR accuracy with sensors on different body locations

Table 2. key stated phrases and quantity statistics from the semi-structured interviews.

Key stated phrases	Number
Needed to conduct data collection involving human participants	10
Gestures were easily deformed, especially when participants felt bored and tired	10
Difficult to label the data	10
Data collection is the most difficult part of HAR research	10
Dull, time-consuming, stressed	10
Data collection were delayed due to COVID-19	10
Would be significantly more productive in doing research if data collection was easier	10
Had to collect the same datasets twice or even more times since they were error-prone	9
Difficult to organize many participants for data collection	8
Unconscious actions during data collection significantly pollute the datasets	8
Felt the tense vibe during the data collection procedure	7
Participants' personalities significantly affected the datasets' quality	7
Felt terrible to beg friends to collect data	6
Participants quit the experiments partway through	6

3.1 Survey

We conducted a survey questionnaire from 50 student researchers who have had on-body sensor-based HAR data collection experience. These researchers are from many countries, such as United States, China, India, Bangladesh and Iran. Their research interests are ubiquitous computing, mobile sensing, gesture recognition, human-computer interaction. Their research work covers a wide variety of applications in HAR, such as mid-air finger-level writing, finger gestures recognition, hand-to-hand gestures, sign language translation, hand gestures recognition for the military, in-air writing for authentication, sleep monitoring, fitness tracking, and investigating HAR accuracy with different sensors and on various body locations. The key questions in the questionnaire included "How much time did you spend on data collection?"; "Did you find that data collection was difficult?"; "What were the difficulties of data collection?"; "What were the potential factors that may affect the quality of the data set?"; "Did you become frustrated in performing the data collection, and if so, can you describe your

frustrations?"; "Did you, at times, feel overwhelmed and temporarily stop or even give up on the data collection?"; "What were the difficulties encountered with the participants?"; "How did you solve the difficulties in data collection and the various problems encountered with participants?";

The following shows the main results. 47% of student researchers needed 2-4 months to collect data, and 20% of them even spent 5-8 months on data collection. 100% of student researchers said that gathering data for their HAR data collections was extremely difficult. Specifically, 79% of researchers thought that data collection was tedious and difficult to sustain; 85% of them reflected that it was difficult to recruit a large number of participants; To our surprise, it shows 89% of collected datasets had poor quality and needed to be collected again. Regarding the reasons for the low quality, 84% of respondents believed that participants' impatience caused it; 89% believed that it was because participants were tired from participating in long periods of data collection; 89% thought that participants were not able to know whether their activities or gestures were appropriate and qualified, which resulted in low quality datasets.

Furthermore, student researchers also thought that there were many difficulties encountered with the participants themselves. Of our respondents 74% of them believed that participants were not serious about collecting data; 89% of them thought that participants felt bored; 65% of them felt like participants showed reluctance; 37% of them mentioned that participants did not show up on time. Based on their experiences, 90% of student researchers reported that they became frustrated while performing the data collection; 63% of them felt overwhelmed, some temporarily stopped or even gave up on data collection entirely. When we asked about how they overcome their difficulties, we were shocked to find that 70% of researchers chose to just push on, even if they were physically and mentally exhausted; 76% tried their best to encourage participants during experiments; 69% chose to take a longer time and do it slowly. Based on the survey results we believe that student researchers do not have good strategies for data collection, even though they had many difficulties in HAR data collection. It is essential for us to explore ways or guidelines to ease the burden of HAR large data collections.

3.2 Individual Semi-Structured Interviews

In order to further understand current practices in on-body sensor-based HAR data collection, we interviewed ten students out of the 50 student researchers. As shown in table 1, they are from different universities, including seven Ph.D. students, two masters students, and one undergraduate student. Three participants were female and seven were male. The video interviews were conducted online and lasted approximately an hour, with a payment of 15 dollars per person. The interview themes included human activity applications, possible factors impacting dataset quality, the time investment required for data collection, and other difficulties collecting data. Key questions included "what kind of human activity or gestures are you trying to recognize?"; "How much time did you used for data collection?"; "what are the difficulties for data collection?"; "what are the potential factors that may affect the quality of the datasets?". To make interviewees feel safe and comfortable with talking about their personal experiences and opinions, we did not record the video interviews. We analyzed the statistical frequency of key phrases and have summarized them in the following paragraphs. Table 2 shows how many interviewees stated each of the identified key phrases during their interviews; the phrases were extracted from the interviews conducted regarding the difficulties for on-body sensor-based HAR data collection and the frequent end result of extremely limited, often inaccessible, datasets.

3.2.1 Difficulties in Conducting Human-Involved Experiments. All the interviewees said that it was difficult to collect on-body sensor-based human activities and gestures. There are limitless amounts of readily accessible data via the internet for several other domains previously mentioned above (speech recognition, natural language processing, and computer vision). In these other areas it is possible to label the data manually since humans can understand data through either listening, reading or watching a video. In contrast, our preliminary study revealed significant challenges in labeling the on-body sensor-based human activity manually after the data collection,

and an additional complication is that participants were usually asked to do specific activities or gestures, using a text transcription for guidance on what kind of activities needed to be performed and in what order. For example, for an IMU finger writing recognition system participants were repeatedly asked to write specific numbers and letters according to the text transcriptions, after which researchers labeled the finger writing data based on the text transcription.

This process was extraordinarily difficult, according to the interviewees' responses. Firstly, it was challenging to organize a group of participants to come to the lab and do the experiments. Then, since the researcher needed to guide and supervise the participant to conduct activities and operate the devices for data collection, participants had to come to the lab at different times. However, this process is very time consuming, organizing and conducting these studies requires significant flexibility on the researcher's part, as well as a large time commitment; during the Covid-19 pandemic this required additional time above and beyond the norm, to ensure disinfecting routines could be followed once in-person research began again. For example, sometimes the devices might crash, meaning that the participant had to conduct the experiment again, which was time-consuming and affects other participants' schedules; additionally some participants did not show at the appointed time, which affected the researcher and other participant's schedules. Secondly, the procedure of the experiment was difficult to endure; performing some specific gestures repeatedly or writing the same graphemes over and over again was extremely boring. Interviewees expressed that the atmosphere during data collection had an intense vibe, many participants had annoyed faces while posing the same activities over and over again. Six interviewees said that they had experienced that some participants did not persist through the tedious process and quit during the experiments. Additionally, all students stated that the COVID-19 pandemic made this data collection more difficult since researchers could not meet the participants in person.

3.2.2 Low Quality of Collected Datasets . Student researchers need to get ground truth training data and so this data must be correct (and labeled as such). Although all interviewees dedicated a lot of time to supervising and monitoring the whole data collection procedure, the collected datasets were usually low quality. One issue is that participants lacked any sensing feedback when they posed a specific gesture, they could not see or hear what gestures they were posing. In this way, their actions were easily deformed, especially when they felt bored and tired. Second, participants usually exhibited unconscious actions during data collection, which significantly polluted the datasets. For example, participants who were required to wear a smartwatch to write numbers in the air needed to touch the "start" button before writing and touch the "stop" button after writing on the touch screen. These two touching actions were required to be completed with the hand without a smartwatch while keeping the hand with the watch motionless since the touching gesture was not in the text transcription for this experiment. However, participants might not hold the hand with the smartwatch stationary when they tried to touch the smartwatch buttons with another hand. Also, participants might put down their hands for resting during writing when they felt tired, and the put-down action would be falsely labeled as an activity in the transcription. Third, since repeatedly doing the same activities for a long time was boring, participants' personalities also significantly affected the datasets' quality. To give an example, in the finger writing data collection impatient participants tended to have sloppy writing, especially when they got bored and tired.

3.2.3 Data Collection Was Difficult for Student Researchers. Many student researchers in the area of on-body sensor-based HAR need to conduct data collection involving human participants. All the interviewees identified data collection as the most difficult part of the research study, stating that it was incredibly dull and time-consuming. This made the student researchers overly stressed since they already needed to put significant effort into course study, homework, exams, algorithm design, and technical implementations. Interviewees often had to resort begging their friends to participate in data collection, and they said that leveraging their friendships like that made them feel absolutely wretched. One interviewee (P8) said that the quality of the data from her friends and non-friends had significantly different accuracy ratings, so she had to beg even more of her friends to

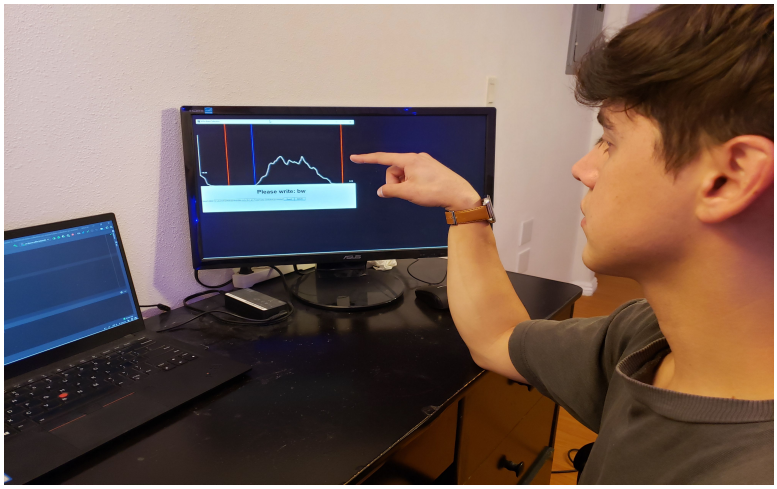

Fig. 3. The dataset collection example: mid-air micro finger writing recognition

participate in order to get higher accuracy results. Eight interviewees emphasized that, in spite of their best efforts, the final on-body sensor-based HAR data collection was usually of low quality. It was difficult to tell if participants performed the activities well or not because the data they got from sensors look like random signal values that researchers could not easily understand. Therefore, they often had to collect the dataset two or more times as collected datasets frequently had abnormally low recognition accuracy, making data collection exponentially harder. All interviewees agreed that if they can save time in the data collection they could be significantly more productive in doing research. As a result, we believe that it is crucial to explore how to overcome these difficulties and help the student researcher collect on-body sensor-based human activity data.

4 DESIGN AND COLLECTION OF A DATASET EXAMPLE

To further understand the difficulties of this type of data collection and to facilitate researchers/students in collecting data, we designed a dataset collection example which we then used to systematically evaluate the key factors that impede data collection. We designed a dataset collection example that is challenging and that meets the following requirements:

1 The dataset includes dozens of participants' data so that it matches the difficulties the preliminary study had, of organizing many participants to join in the data collection.
2 Each participant is required to perform activities for a few hours so that each participant will feel bored due to posing the same gestures again and again for hours.
3 The dataset should not be camera data or microphone data which have visual or acoustic feedback. Thus, it is difficult for participants and researchers to know whether they perform the gestures well or not.
4 It should be easy to unintentionally produce unconscious action noise data into the collection procedure.

To meet all these requirements, we designed the data collection experiment to be a mid-air micro finger writing data collection with IMU sensors. Mid-air finger writing recognition needs to label 36 classes of graphemes (ten numbers and 26 letters). Note that although the smartwatch is not moving much, or at all, the moving finger causes vibrations which can be captured by the IMU sensor. This fine-grained finger writing recognition requires hours of training data [11], which is difficult to collect since the finger writing vibrations are not something people, even researchers, intuitively understand. Additionally, it is easy to make unconscious motions between

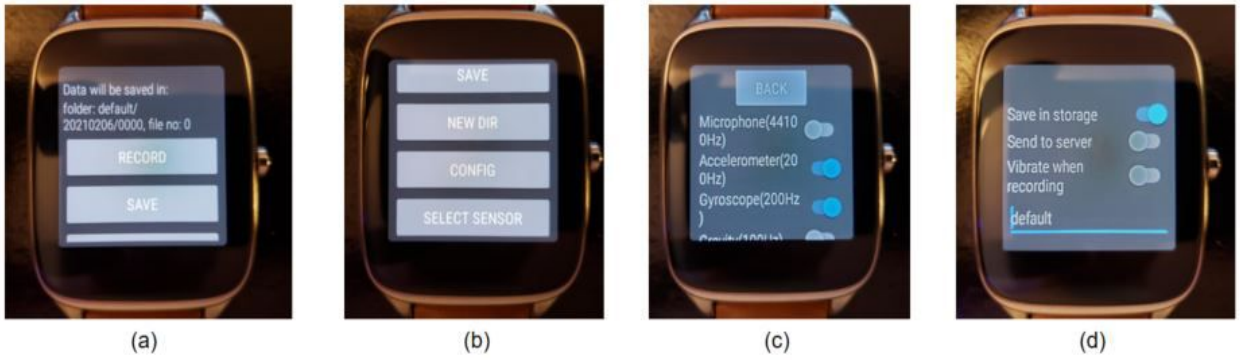

Fig. 4. An app to collect sensor data for Android devices

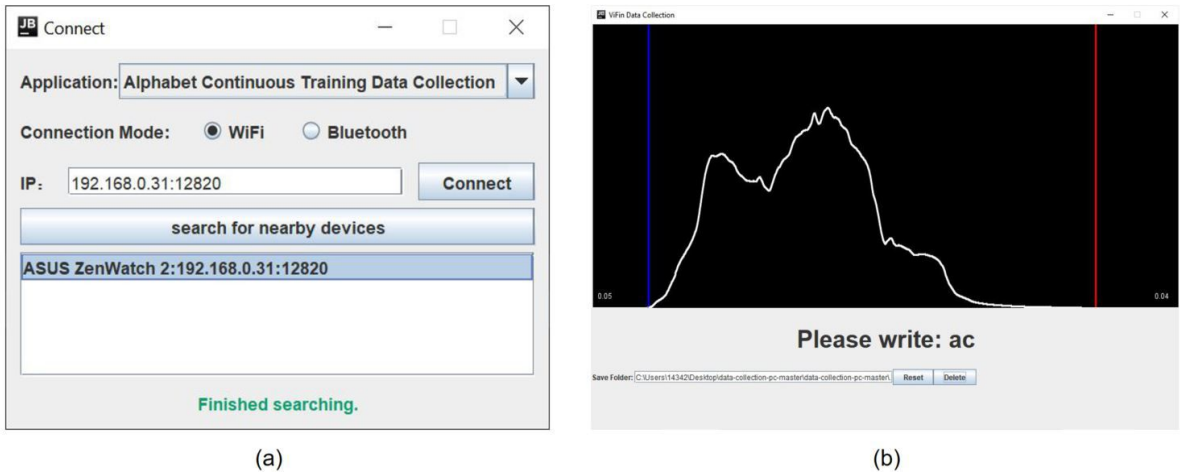

Fig. 5. VFC software for Windows and Mac to provide real-time sensor data output

writing graphemes, which can cause erroneous labelling and pollute the data quality.. Last but not least, different users have different writing styles, thus many participants' data is necessary for testing how well the training model works across different participants. Therefore, this dataset's collection criteria is comprehensive enough to match all the challenges mentioned in the preliminary study.

For this data collection study, we required participants to wear a smartwatch and move their index fingers to write numbers and letters in the air as shown in figure 3. [11] Each participant was required to write for six hours. We used this dataset to train models for continuous mid-air finger writing recognition. We developed the data collection system on Windows, Mac and Android devices to collect finger writing data. The data collection system and the dataset we collected are available at https://www.jianguoyun.com/p/DVaDq-QQ2enUCBiLx-AD for further research.

4.1 System Implementation

We developed an app, named Sensor Data Collection, on Android devices. With this app, users can collect sensor data from mobile devices. As figure 4 (a) shows, users are able to click on the "record" button to collect data from the sensors in a smartwatch. After clicking on the "record" button, it turns into the "stop" button. Users can click "SELECT SENSOR" in the figure 4 (b), and all the sensors in the android device show up in a list as shown in figure 4 (c). In this case, we choose "accelerometer" and "gyroscope" to detect the finger writing. Users can click on "CONFIG" in the figure 4 (b), then choose to save data in the smartwatch (save in storage) or in the laptop (send to server). We also developed a software, named VFC, for computers with Windows or Mac operating systems. Participants are required to connect a smartwatch and a laptop on the same WiFi through the app "Wear OS". Then participants choose the WIFI option and click the button "search for nearby devices" to connect the smartwatch as shown in figure 5 (a). After that, our software shows real-time IMU signals on the computer screen so that participants are able to see the signal outputs when they move their fingers. Figure 5 (b) shows a signal wave of finger writing. VFC uses a network socket [2] to send IMU data from the smartwatch to the laptop. After we collected data, we trained the model in TensorFlow using a unidirectional Gated Recurrent Units (GRU) with an extra connectionist temporal classification (CTC) layer. This model has been used in previous research works. [10] GRU is an efficient Recurrent Neural Network that has memories, but has fewer gates (reset and update gates) than Long Short-Term Memory (LSTM). CTC aligns the input and output with a token "∅" and computes the loss much faster using the Forward-backward algorithm than when using the method of exhaustion. To be specific, the GRU has one hidden layer and 32 neurons. We choose Adam to optimize network training. The sliding window size (time frame) is 0.5 seconds and the moving step is 0.1 seconds. Since IMU has six axes of data, we concatenated each window's data of six axes in series as a single input to the GRU.

4.2 The Example Dataset

In this section we present our datasets. The training set includes numbers and letters for continuous finger writing recognition. In order to validate that the training set is qualified (labeled correctly), we designed a validation set. Note that to evaluate if the qualified training set works in the real world, we also conducted a user study to evaluate the real-time finger writing recognition in practice (see section 6.4).

4.2.1 Training Set: We collect numbers (0∼9) and letters (a∼z) to recognize continuous finger writing with an IMU. In order to learn every transition between graphemes, we collect each pair of numbers (00, 01, ..., 98, 99) and each pair of letters (aa, ab, ac, ...za, zb, zc,..., zz) for training, which results in 100 pairs of numbers and 676 pairs of letters. Participants are required to repeat the training set collection five times (rounds). In total, there are 3880 pairs (776×5) of graphemes in the training set for each person, which is six hours' data in total per person.

4.2.2 Validation Set. To determine if the training set is accurate and qualified, we collected additional validation data. Since our dataset is designed to recognize continuous finger writing, the validation set includes numbers and English sentences, which participants write continuously. To evaluate the accuracy of recognizing numbers, 100 random numbers are collected for each person. These 100 numbers are split into 10 groups and participants are required to write them continuously in each group. The random 100 numbers are grouped as follow: [1,3,2,4,6,8,8,9,1,3]- [2,0,5,4,7,4,8,5,9,7]- [9,9,7,1,4,0,5,3,3,0]- [7,1,8,1,3,3,3,0,2,5]- [8,7,9,6,9,5,2,9,7,4]- [8,9,5,8,7,5,5,2,3,6]- [2,1,6,1,6,8,5,2,0,6]- [2,4,8,4,4,0,6,7,2,0]- [6,0,4,4,5,6,3,1,7,9]- [3,6,2,1,7,9,0,8,1,0]. Note that some numbers are repeated in the same line. In order to balance each letter for evaluating the accuracy of each letter, we collect seven English sentences, where each sentence contains 26 different letters. The first sentence is 26 letters in order from a to z. Participants are required to write them continuously. The rest of the six sentences are called perfect pangrams [3] and consist of some English words. In perfect pangrams, each letter of the alphabet occurs once and only once. Between different words, participants are required to pause according to their writing habits.

Table 3. Organization of Participants.

Group	Investigators	Participants
A	2 students in our research group working on this mid-air finger writing project	Each student collects the number of participants' data as many as they can
B	8 lab mates not working on this finger writing project	Each lab mate collects her own data (one participant)
C	18 random undergraduate students	Each student collects three participants' data

Table 4. Control variable experiments.

Group	Investigators	Expected participants	Variable
1	3	9	w/o Sensing feedback
2	3	9	w/o Music or videos
3	3	9	w/o Data quality requirement
4	3	9+	w/ Monitoring
5	3	9	w/o Collection period requirement
6	3	9+	SenseCollect: w/ sensing feedback, w/ music or videos, w/ data quality requirement, w/o monitoring, w/ collection period requirement
Note: Investigators are the people we recruit and interact with, while participants are recruited by Investigators to write numbers/letters with IMU sensors.			

[abcdefghijklmnopqrstuvwxyz]- [mr jock tv quiz phd bags few lynx]- [cwm fjord bank glyphs vext quiz]- [blowzy night frumps vexd jack q]- [squdgy fez blank jimp crwth vox]- [tv quiz drag nymphs blew jfk cox]- [q kelt vug dwarf combs jynx phiz].

4.3 Collection Procedure

The goals are to collect a large and qualified on-body sensor-based human activities dataset and find an efficient way to achieve it quickly. We design a series of control variable experiments [1] and analyze the challenges mentioned by interviewees in the preliminary study. To explore and find an efficient way to collect this dataset, we try three different methods to organize participants for performing finger writing as shown in table 3. To further analyze what potential factors may complicate data collection, we use the control variables method to investigate how the following factors affect the data collection: sensing feedback, psychological feeling, data quality requirements, monitoring, and the data collection period. For each factor, we recruit three students to do the experiments. These students are expected to collect data from 54 participants in total. Table 4 shows the six groups with different control variables. Note that all the experiments involving human subjects conformed to the relevant regulations of our institute.

4.3.1 Organization of Participants. We investigate how to organize participants to collect data. We conduct three sets of experiments on organization outlined in table 3. First, for (Group A), we ask two students in our

research group working on this mid-air finger writing project to manage and collect data directly. (According to the previous survey, most of the student researchers directly manege and collect data.) We require them to collect as much data as they can in two weeks. Since these two students are the research students working on the finger writing recognition project, they do not get extra payment for data collection. Second for (Group B), we request eight lab mates (participants) to help to collect data. They are not working on this finger writing recognition project. We committed to buying a pizza lunch for everyone after the data collection as an incentive. We make a guide file and video to help them on setting up devices and collecting their finger movement data. Third for (Group C), we recruit 18 random undergraduate students investigators to manage and collect data with different variables as shown in table 4. We provide the same file and video to guide them to collect data. Each sub group has three investigators. Each investigators is required to collect three participants' data in a week. Thus, we expect 54+ (6*3*3) participants' data to be collected. Note that we expect more than 9 participants' data to be collected in group 4 and 6. (see section 4.3.2) Here "investigators" are the student we recruit and interact with, while "participants" are the people who directly wrote numbers/letters with IMU sensors. Participants can be anyone, such as investigators' family or friends. After the investigators complete the data collection, we ask them some questions for feedback. For example, how did the data collection process go? Are you willing to collect more participants' data? Are you willing to do some similar part-time jobs for data collection in the future? The purpose of these questions is to understand the investigators' data collection procedure and investigators' personal feeling for a data collection part-time job. We pay these undergraduate investigators 450 Chinese Yuan (70 US dollars) for a participants' data (six hours). We did not dictate how much the students pay the participants (More details about this see section 7.1). Since group C requires many participants for different comparison experiments, we recruit group C in China to stay within budget.

4.3.2 Control Variable Experiments. In this subsection, the design of the control variable experiments for Group C, that explore the potential factors affecting data collection difficulty, are shown in table 4. Within Group C, Group 1 does not get the VFC to provide sensing feedback for them. Group 2 does not listen to music or watch videos. Group 3 is told they would only get paid when the data quality meet the requirement. The students in Group 4 are required to supervise and monitor the participants' whole data collection procedure. We require Group 5 to complete a participant's data collection without a data collection period requirement. Group 6 is the control group, which has the VFC system to provide sensing feedback and music or video during finger writing. Group participants are told that only get paid if the data quality meets the requirement, and they do not require monitoring, and the participants cannot collect data for more than two hours in one day. We will discuss the details of each control variable in group C in the following paragraphs.

Sensing Feedback. We developed a software, named VFC, to show real-time IMU signals on the computer screen so that participants are able to see the signal outputs when they move their fingers. In order to prevent participants' unconscious movement, we also detect the start point (blue line) and endpoint (red line) of the finger writing as shown in figure 5 (b). For example, When VFC notice participants write a pair of letters of "ab," users move their index fingers to write in the air. The IMU sensor data is sent to the laptop and shown on the screen as shown in the figure 5 (b). VFC calculate the overall energy of six axes of IMU then use the energy-based threshold approach [9, 13] to detect the start point and the endpoint. (We set the threshold as 0.35 here). When the red line show up, the notification asks participants to write the next pair of letters, which is "ac" in the figure 5 (b). Then participants should start to write again. In this way, if an unconscious movement is detected and shows up on the screen, participants can click the "delete" button, so that the unconscious movement does not be annotated as letters writing. For group 1, we do not offer this software for them to collect data, but ask them to write according to a text transcription. In contrast, we give this software to the rest of the students, including participants in groups A and B.

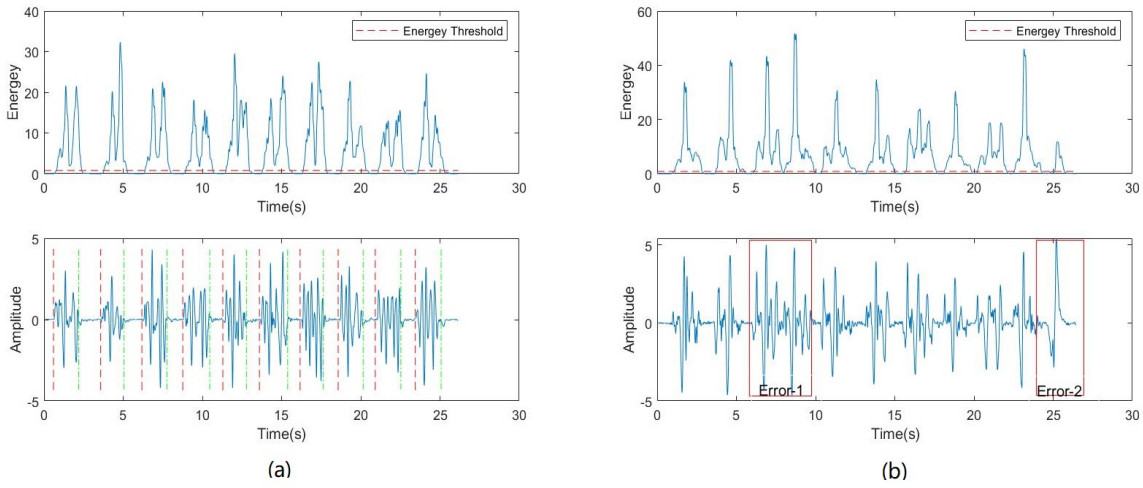

Fig. 6. The visual inspection phase for training data. Numbers: 80,81,82,83,84,85,86,87,88,89.

Psychological Uninteresting Nature. The psychological uninteresting nature of data collection is one of the reasons that cause on-body sensor-based HAR data collection to be very difficult. Writing simple numbers and letters for hours is tedious and causes mental fatigue. This reminds us people working in a factory also have trouble focusing on the tedious and straightforward works. Music in factories [27] shows that music affects industrial workers' feeling and help product quality control. Therefore, for three undergraduate students in group C, we do not have them provide earphones for participants to listen to music or watch videos during writing in the air with the index finger, while we highly recommended the other groups ask participants to listen to the music or watch videos on the laptop during data collection. We investigate how the music and video may affect participants' feelings of boredom and the datasets' quality.

Data Quality Requirement. Data quality requirements for the participants may affect the collected data quality. We do not pay the participants who collect unqualified data so that they are motivated to collect data carefully. However, we do not tell the investigators and participants in group 3 about this requirement. We want to see whether the data quality requirement will affect the data quality or not. To evaluate the data quality, we first have a visual inspection phase to check whether the collected data is qualified. For example, there are ten pulses for collecting number writing (80, 81, 82...89) in the training set. We segment these ten pulses and label them 80, 81, 82...89 in order, as shown in figure 6 (a). The red line is at the start point of a signal and the green line is at the end point of the signal for labeling. If there are more or less than ten pulses, we will ask the investigators to re-collect the writing. For example, as shown in figure 6 (b), the error-1 shows the numbers 82 and 83. The participant did not pause between these two writing, which will make the system label them as number 82. And the rest of the data will all be mislabeled. Further, error-2 in figure 6 (b) shows there is an extra gesture at the end. We believe that it was because the participant did not hold the hand with the smartwatch stationary when they tried to touch the smartwatch "stop" button with another hand. This extra gesture will be mislabeled as a number writing. Further, participants may write numbers/letters sloppily. In addition, participants may write a wrong letter/number but do not correct it, which we will not be noticed through a visual inspection. These poor quality data can not be found through a visual inspection since the collected data is just some random number sequences

no matter whether it is an unintended extra movement, sloppy writing, wrong writing, or correct writing. Thus, the visual inspection (segmentation and counting numbers of the pulse) is not enough to secure the data quality.

As a result, we use recognition accuracy to further evaluate the quality of the collected datasets. To estimate how much recognition accuracy should be considered as qualified data, three of us in this project collected six-hours of training data (five rounds) for training individual models (see section 4.1). The lowest accuracy of three person was at 90% for five rounds of dataset, and at 50% for each round of data. This is because more data improves the recognition accuracy and one round of training data is not enough to get a good recognition model. And we make a requirement for the participants that we only pay them if the recognition accuracy for five rounds of data is above 80% and the accuracy for each round of data is above 40%. Note that the recognition accuracy here is also one indicator that we use to evaluate the data Key Performance Indicators(KPIs) under the ablation study with different control variables. In order to study the influence of different variables, we also collect a validation set in the same way as the training set with different controlled variables. Note that here we are trying to collect qualified training datasets and label them correctly, not a qualified test set. For example, if an unintended movement is collected and labeled as writing, the wrong labels will pollute the datasets. Note that we later evaluate the accuracy of the collected datasets with a real-time system. (see section 6.4) The system can recognize the numbers/letters in the real world when the participants write a standard transcription in natural settings, which may include unintended movements.

Monitoring. Supervision and monitoring are frequently used during data collection. All the interviewees in the preliminary study told us they monitored the whole data collection procedure of participants to ensure the participants were doing it correctly and there was no problem during data collection. However, watching the entire data collection procedure is time-consuming. This section investigate the necessity for supervising and monitoring the participants the whole operation while they perform finger writing. In group 4, we ask three undergraduate students to supervise and monitor the participants during the whole process when participants move their fingers to write numbers/letters for data collection. The rest of the students in other groups only guide participants on collecting the data at the beginning (no monitoring for the whole procedure). Also, We want to know how the monitoring requirement affects the number of participants' data which students are able to collect. Thus, we ask the students in this group (group 4 with monitoring) and the control group (group 6 without monitoring) to collect as many participants' data as they can.

Data Collection Period. Performing monotonous and straightforward tasks, again and again, causes mental fatigue. Thus, the quality of datasets may get affected if participants are collecting data for a long time. For example, participants may start to write sloppily. In this experiment, we ask three students in group 6 to collect six-hours of data from one participant on the same day. Participants are allowed to take a break when they feel tired. For the rest of the students in other groups, we ask them to not collect data from the same participant more than three hours a day. We investigate how the collection period will affect the recognition accuracy.

5 DEPLOYMENT STUDY

This section shows our findings from the deployment study and presents SenseCollect guidelines as shown in table 6.

5.1 **It was difficult for research students to collect data by themselves.**

Although the two students in group A were directly responsible for the finger writing recognition project, they only collected 2 and 3 participants' data in two weeks, respectively. The students claimed that it was tough to collect the data. First, the data collection was time-consuming and the students had other responsibilities (i.e. classes, exams, and other research projects), which slowed down their data collection. Second, one of the students

Table 5. Results of control variable experiments.

Group	Variables	Quit investigators	Participants qualified	Participants unqualified	Accuracy qualified	Accuracy unqualified
1	w/o Sensing feedback	2	0	3		31%±13.6
2	w/o Music or videos	1	5	0	89%±6.1	
3	w/o Data quality requirement	0	0	9		62%±15.5
4	w/ Monitoring	0	6	0	91%±5.9	
5	w/o Collection period requirement	0	4	5	85%±4.3	74%±4.1
6	SenseCollect	0	12	0	90%±6.3	

Table 6. SenseCollect Guidelines

1. Recruiting investigators to collect data.
2. Sensing feedback is essential.
3. Music and videos facilitate the collection procedure.
4. Data quality requirement is important.
5. Monitoring is not necessary.
6. Collection period matters.

claimed that the participants did not want to conduct the finger writing activity for 6 hours. He begged many of his friends, but most of them rejected him. Although a few people would like to help with this experiment, some participants quit after a few hours of finger writing because they were bored and tired. Third, after six hours of data collection, sometimes the data accuracy was extremely low (lower than 30%), but the participants rejected to redo the experiments. Regarding data accuracy, only two participants' data met the quality requirement, while the rest of the data had a low accuracy at 58% on average.

Regarding the data from the eight lab mates (Group B), only one participant's data was labeled and ran through the algorithm. However, the accuracy was only 22%. We were not able to label the other seven lab mates' data because their data was insufficient. For example, for writing numbers 0 to 9, there should be ten pulses in the signal while there were fewer or more pulses. Additionally, the number of the files did not match the text transcription we provided, so that we could not label them for model training.

By recruiting undergraduate students to collect data in group C, we successfully collected a larger dataset in a short time (one week) in the group 6 where people following the SenseCollect guideline. Three investigators collected 12 participants' data in one week and all the data is qualified with an average accuracy of 90% and with a standard deviation at 6.3. (4 participants/investigators/week on average).) Note that two researcher students (Group A) only collected two qualified participants' data in two weeks (0.5 participants/student/week on average). And eight lab mates (Group B) collected no qualified data.

In conclusion, it was difficult for graduate students to collect data by themselves, and based on these findings we recommend recruiting undergraduate investigators to collect data. We will further analyze the affecting factors in group C with a control variable study in the following paragraphs. The results are shown in table 5. Surprisingly, investigators collected data faster during the COVID-19 pandemic when the school was in lock-down. This could

be a result of students collecting data from their families who lived with them during the pandemic, making easier to recruit participants. Students might also have had more time during the pandemic than normal times.

5.2 Sensing feedback is essential.

The three investigators in group C who did not get the VFC complained that collecting finger writing data was very difficult to complete. First, they gave the text transcription to the participants to write; however, participants usually missed some parts in the text transcription. To combat this problem, the three investigators had to read the text transcription out loud for the participants to write. Reading the text transcription out loud for six hours was exhausting for the investigators, and the collected data still sometimes missed parts, according to the text transcription, or the collected data files had more signals than there should be. For example, for numbers 0 to 9, it should have ten impulses in the file, but the collected file had more than ten impulses. We believe that it was because participants had some unconscious actions during writing. As a result we were not able to label the data based on the text transcription, and we had to ask the investigators to redo their experiments if their collected data did not match the text transcription. As a result, two of the investigators quit the job, as the table 5 shows. One investigator collected three participants' data, but the accuracy is only 31% with a standard deviation of 13.6. We believed that even though the collected data matched the text transcription, the participants might not write the right numbers/letters they are supposed to write. If it was writing numbers/letters with a pen on paper, it was easy to know whether participants wrote correct numbers/letters or not. However, we could not recognize the writing data of IMU, so without the sensing feedback participants did not know how well they were writing. In contrast, for the students who used the VFC, it is easy for participants to follow the application to write. As figure 5 (b) shown, the VFC will show up a pair of numbers/letters for users to write. At the same time, users will see the real-time signal output from the IMU sensor to see if they made unconscious actions. If so, they can click the "delete" button on the VFC and rewrite it. Thus, the accuracy of the control group (SenseCollect) shown in table 5 is much better (90% in average).

5.3 Music and videos facilitate the collection procedure.

From the interview in the preliminary study, a psychological uninteresting task was one of the biggest problems. It was difficult for participants to write numbers/letters repeatedly for hours. Research has shown that individuals are generally more engaged in activities they enjoy or consider valuable, which leads to better performance [33], it is reasonable to conjecture that individuals engaged in tedious data collection they dislike or don't value will have poorer performance. We suggested participants listen to music or watch videos with earphones to mitigate boredom. For the three investigators without this suggestion, one investigator quit the job. He told us that some participants wrote the first hour but refused to come back to continue. He believed this part-time job was too complicated and not worth it to him for the payment we offer. Although the other two investigators collected five participants' data, they refused to collect more after that. Also, they said that this job was challenging and they would be busy in the following month so that they would not able to do this part-time job in the near future. For the participants who were notified that they could choose any music to listen to or videos to watch, all the three investigators completed the job and said they would be happy to do a similar job again in the future. Therefore, we believe that music and videos made participants feel better and helped them perform simple and boring data collection.

5.4 Data quality requirement is important.

We did not have a data quality requirement for three students in group 3. We told them as long as the number of data matches the text transcription, we would make the payment. Although all the students completed the job, the data we got from them had a low accuracy rate, only 62% on average. We believe that participants were

motivated to complete this tedious data collection procedure as fast as possible, which resulted in sloppy writing; additionally they were unlikely to delete and correct their mistakes when they wrote numbers or letters wrong. Since this resulted in a low-quality training set of data we took a different approach with later groups. The concept of Pay for Performance (PFP) is used in organizational psychology to improve employee work quality by having their performance influence their pay. [21] We thought this might work in data collection as a way to motivate participants to improve their performance on the tasks. To our surprise all of the participants in group 6 met the data quality requirement, with an average accuracy of 90% overall. This suggests that the concept of PFP was useful in improving data quality for HAR experiments, both because participants were motivated by the extrinsic reward of payment, and because they did not want to waste their time doing a poor job and then not getting paid, so they did the experiments carefully. Also, some people (including investigators and participants) immediately left when they found out they may not get paid if the collected data is not qualified. We believe that PFP was also helpful in filtering out the people who were unable or unwilling to perform this job well. We asked the investigators what the reactions of participants were when participants knew that they had a risk of not getting paid, and whether the participants felt stressed during experiments. The investigators told us that the people who took the job means they were willing to take the risk, and people who were unwilling to take that risk left instead of participating. Also, investigators said it was helpful that we gave them their model accuracy as feedback to them on each hour's data, especially for the first hour's data, as this made the participants believe that they would get the payment as long as they kept doing seriously.

5.5 **Monitoring is not necessary.**

In the beginning, we believed that asking the students to supervise and monitor the participants to do the data collection would make participants write carefully, resulting in improved data quality. Thus, we asked three investigators to watch the whole data collection procedure (six hours per participant). We did not require monitoring for the rest of the investigators aside from a tutorial done at the beginning of data collection. To our surprise, there were no differences in accuracy between these two groups. However the investigators in group 4 only collected six participants' data, while the investigators who did not monitor the data collection procedure in group 6 collected twelve participants' data, which was twice as much than in group 4. Investigators complained that monitoring was tedious. We believed that asking for monitoring for the entire procedure was time-consuming and caused difficulties for investigators, making them only obtain a small number of participants' data.

5.6 **Collection period matters.**

We required three investigators to collect a participant's data without the requirement of collection period in group 5 while asking the rest of the students not to collect a participant's data for more than two hours in one day. We found out this affected the quality of the dataset. In group 5, although all the students completed the data collection (9 participants' data), only four participants' data met the data quality requirements with an average accuracy of 85% (4.3 standard deviations). Note that this was lower than the accuracy of 90% in the control group, which we required them not to collect too much data on the same day from one participant. Furthermore, five participants' data did not meet the data quality requirement (at 74% accuracy on average). Participants tended to finish the burdensome work in one go based on our observation, making them write sloppy when they were tired. Note that although we have the same data quality requirement for group 5, we still paid them for the unqualified data (five participants) in the end.

5.7 SenseCollect Theory

Based on the above analysis we determined that it is more efficient for the researcher to recruit other students (investigators) to manage and collect data from participants rather than directly collecting data from a large amount of participants. Unfortunately, based on our survey results all of the student researchers are directly collecting data from a large number of participants, which is not only physically but also mentally stressful. Remember that 90% of student researchers reported that they became frustrated while performing the data collection, with the survey showing high frustration levels (avg 8.6/10 points). One reason why the student researchers preferred to directly collect the data from participants is that they believed they had to monitor the participants in order to secure the HAR dataset quality; based on the survey and interviews 90% of students directly monitor the participants during the data collection process. However, based on our findings monitoring HAR data collection is not significantly helpful; in fact directly monitoring is potentially harmful, something the student researchers would not have known. Monitoring is not only time-consuming, which negatively impacts the researcher's efficiency, but based on our findings it is not helpful in securing the dataset quality. SenseCollect shows that, for the variables we tested, implementing a data quality requirement for payments is the key to securing the dataset quality, not the traditional monitoring strategy which is currently used by most student researchers. Research [21] shows that pay for performance (PFP) models improve employees' work quality. Similarly, our study found that the participants in the payment for performance group collected higher quality data than participants in the payment for participating groups. Additionally, the surveys and interviews also indicated that 89% participants were unable to know whether their activities or gestures were appropriate and qualified. This emphasizes the need for real-time sensing feedback in HAR studies; SenseCollect transmits and vitalized the collected data on screens in real-time. This does require student researchers to develop systems for transmitting and vitalizing real-time sensor data on screens (e.g., laptop, phone, smart glasses), but providing sensing feedback can significantly improve data quality. Another reason for poor data quality is that during the data collection participants become both physically and mentally tired. Typically tired humans make more mistakes, and that means participant fatigue will detrimentally impact the participant's dataset, impacting the overall quality of the data [18]. However, despite finding the work burdensome we observed that participants preferred to complete the work in a single long session, rather than splitting the work up into multiple shorter sessions. The average accuracy of datasets for the first hour and the second hour is 58% and 60% respectively, but from the third hour through the sixth hour the accuracy of each hour's data dropped to 40% or lower. Given the significant drop in accuracy it is vital for student researchers to remind, or even require, the participants to take a break during the data collection to reduce fatigue-related data pollution. Additionally, motivation has also been shown to mitigate the negative impact of fatigue on performance; [6] therefore by utilizing a PFP structure SenseCollect creates significant motivation for participants to achieve high data accuracy, which should both increase participant willingness to take a break, and help offset the fatigue-induced detrimental impact on performance. Currently, to improve engagement and reduce boredom and mental fatigue many student researchers (76%) reported that they tried to encourage participants during the experiments; this was laborious for the researchers and had little to no positive impact on participants. SenseCollect shows that playing music or videos during tedious experiments eases the mentally boring tasks and improves the engagement of participants. Although music in factories is proven helpful to engagement in existing publications [27] and is widely applied in many factories, music is still neglected or doubted by researchers for being used in HAR data collection based on our study results.

To help a future researcher to decide which of the variables they should prioritize, we ranked the different control variables based on their contribution to SenseCollect's performance. Currently most student researchers are still directly monitoring and collecting data, a fact which needs to change if we want to improve HAR data collection. Therefore our top guideline is that researchers need to learn to trust and recruit investigators for

Table 7. Evaluation of collected dataset

Evaluation	Main results
Offline accuracy	90% on average
Size of collected data	32 participants (192 hours of data)
Speed of collection	one week
Real-time system accuracy	85% on average
Possible future research based on this dataset	a general model working across different users
Generalizability of SenseCollect guideline	1. SenseCollect guideline in different HAR domains 2. another dataset collection based on SenseCollect guideline

data collection, to improve the efficiency and quality of data collection. To help the investigators to collect data successfully real-time sensing feedback, which helps the participant know whether their performing activities or gestures are appropriate and qualified, is the most important variable. Without real-time sensing feedback, no qualified data was collected. The second crucial variable is the data quality requirement for payment; without this variable, no qualified data was collected, but the collected data without PFP had better quality (62%) than the quality of collected data without sensing feedback (31%) which is why we ranked sensing feedback first. The third significant variable is the collection period requirement; without this requirement we had some unqualified data but there was some qualified data without this variable, unlike with sensing feedback and PFP. Music playing is ranked fourth because a lack of music did not cause unqualified data but it did make one investigator quit. Lastly, monitoring was harmful instead of helpful, it caused investigators to recruit fewer participants.

6 EVALUATION

In this section, we evaluate the example dataset we collected. We elaborate on the accuracy of the dataset both in offline and in a real-time system. Also, we present the size of collected data and the speed of data collection. Then we discuss the possible research based on the collected dataset. Lastly, we illustrate the generalizability of SenseCollect guideline in different HAR domains. A short summary of SenseCollect evaluation is shown in table 7.

6.1 Offline Accuracy.

To evaluate whether the training set of finger writing data we collected was correctly labeled and qualified, we tested it with the validation set. The classification model is descried in section 4.1 . The recognition accuracy for each individual is about 90% on average. In contrast, the accuracy of data without sensing feedback is only 31%. Further, the accuracy without data quality requirements is only 62%. Without collection period requirements, even the qualified data had a lower accuracy at 85% and the unqualified data had a low accuracy at 74%. Although music and monitoring did not affect the accuracy, it affected the amount of data the investigators could collect. Therefore, SenseCollect is the best solution to collect on-body sensor-based HAR data, comparing to the ways for the other groups.

6.2 The Dataset Size

We collected five rounds of data for each participant. Results shows that the recognition accuracy for each individual with one-round data was only 48% on average. However, the recognition accuracy for five-rounds

of data was 90% on average. This shows that the accuracy was improved significantly with a larger dataset. To our best knowledge, the largest publicity available on-body sensor-based HAR dataset are Daphnet freezing of gait dataset [5] and PAMAP2 dataset [40]. Daphnet freezing of gait dataset [5] only has 5 hours of sensor data in total collected from 10 participants, and PAMAP2 dataset [40] has 7.5 hours of sensor data collected from 9 participants. In contrast, we collected 192 hours of data, which is 38 times and 15 times more than these other datasets, separately. Also, the number of participants (32 participants) is about three times more than them. To be specific, this new dataset includes 27 participants' qualified data from group C, 2 participants' qualified data from group A, and 3 participants data from three of us in this finger writing project. 17 participants' data in group A was not qualified because they were performed with inadequate control variables. Three participants' data in group A and eight participants' data in group B were not qualified.

6.3 Data Collection Speed

In this section, we discuss the speed for collecting qualified data in different ways of organizing participants and with different factors. Unqualified data is not discussed here. In group A, we only collected two participants' data in two weeks (one participant's data in a week). In group B, we did not collect any qualified data. Then we present the speed of collection for six different groups in group C. We required the students in group C to complete the data collection in one week. From group 1 to 5, the collected qualified data was 0,5,0,6,4 participants' data, separately. In contrast, SenseCollect of group 6 collected 12 participants' data in a week, which is much more than the others. We believe that if we recruit more students in group 6, SenseCollect could collect even more qualified data in one week. Thus, SenseCollect is an efficient way to collect large datasets from many participants.

6.4 Recognition Accuracy in Practice

To study the real performance of this dataset in practice, we conducted a user study using a standard text transcription. We recruited ten participants from those included in this dataset. We trained individual models for each participant and put the models into a smartwatch. Participants were required to have a seat, place their elbows on the table, and write the text transcription for two hours. The screen of the laptop shows the text output in real-time. We used a published phrase set for text entry by MacKenzie and Soukoreff [34]. The standard text transcription has 500 phrases. The phases vary from 16 to 43 characters (mean = 28.61). There are 2712 words (1163 unique) varying from 1 to 13 characters (mean = 4.46). The experiment was conducted only with the lower case letters (no upper case letters, punctuation, or numbers).

Text entry speed was measured in words per minute (wpm). We calculated the number of words the participants wrote in two hours. The average text entry speed for ten participants was 15.0 wpm with a standard deviation of 4.2. The average accuracy for two hours was 80% with a standard deviation of 2.7. We also calculated the average accuracy for the first hour at 85% with a standard deviation of 3.3. We believe that the decreased accuracy in the second hour was because participants got tired and did sloppy writing in the second hour.

6.5 Possible Research Based on This Dataset.

We believe there is much future research that could be conducted based on this dataset. For example, when we ask a new user to use the smartwatch for finger writing, the accuracy is inferior. This is because different users have different hand characteristics such a length, amount of muscle and fat, etc. and also write in different fashions. For training a general model working across different users, much research could be conducted based on this dataset, such as transfer learning, data augmentation, domain adaptation, general adversarial networks and so on.

6.6 Generalizability of SenseCollect

We summarize HAR into four domains as shown in figure 1: daily human activities (e.g., walking, sitting), Human-computer interaction (e.g., gestures, writing, and typing), HAR authentication (e.g., gait), HAR related health (e.g., sports). In order to label human activities, all these domains require participants to perform specific gestures/activities one by one based on notifications or text transcriptions. For example, the participants are required to perform sitting for a while, then perform walking, then other activities, in an order based on the notifications or a text transcription. SenseCollect guideline works for all those HAR data collection campaigns. Additionally, we believe that the data collection of gestures (e.g., fitness posture, finger writing) is more challenging than daily activities (e.g., walking and running) since participants need to pay more attention to perform the required gestures well. Non-HAR health data collection, such as monitoring heartbeat by PPG/ECG, is not studied in SenseCollect. However, some SenseCollect guild lines may still be instructive. For example, pay for performance may be helpful to encourage participants to properly place the sensors on their bodies and keep the devices charged all the time.

To further evaluate the efficiency and generalizability of the SenseCollect guideline, we utilized SenseCollect's guidelines to collect another dataset: on-body tapping data. On-body tapping recognition with a smartwatch is a new application for HAR, which recently won the Best Demo Award in Sensys 2020. [12] It uses a smartwatch to detect and localize (classify) passive tapping-induced vibrations in different body locations, thus providing an extended keypad. In our study, we required participants to tap on four left knuckles on the back of the left hand wearing a smartwatch, and each knuckle had to be tapped 30 times. The smartwatch was worn on the wrist in a comfortable manner with the hand floating in the air. The four knuckles were tapped randomly notified by VFC. Following SenseCollect guidelines, we recruited ten student investigators and asked each of them to collect at least ten participants' data over the course of one week. In the end, we got 113 participants' data in one week. The length of the training and test datasets are 20 samples and 10 samples for each key of a person, respectively. Utilizing the neural network presented in the previous research [12], we trained individual models for each participant and got a recognition accuracy of 91% on average with a standard deviation of 8.7. In the interest of improving the availability of HAR datasets, we also made this dataset publicly available at the same link.

7 DISCUSSION

7.1 Payment of Data Collection

We paid the undergraduate students 450 Chinese Yuan (70 US dollars) for a participant's data (six hours). We did not intervene in how much the student pay the participant who wrote numbers and letters with IMU sensors. However, after the students completed the data collection, we asked them how much they paid each participant. The answers varied from 250 Chinese Yuan to 400 Chinese Yuan. We found out the more they paid themselves for each participant's data, the less number of participants' data they got. We believed that it was because 1) it was more difficult to recruit participants with lower payment; 2) students might have got enough payment even from a few participants' data since they paid themselves with a large amount of money. What's more, we recruited students to perform the experiments in the USA at the beginning. However, with the same amount of payment, it was difficult to recruit students in USA. We believed that it was because of participants' financial status. Considering our budget, we decided to recruit group C in China.

7.2 Algorithm, Computing Power and Data

A considerable amount of machine learning algorithms are being designed and proposed every day. Also, in the book of "A Brief History of Humankind" [22], the author mentioned that computing resources, such as high-performance computers and embedded computers, will be the most valuable investment in the future, as like

the valuable investment of pigs in the past and the valuable investment of houses nowadays. Countless researchers contribute to machine learning algorithms and computing hardware nowadays in contrast to fewer researchers paying attention to data collection. We believed machine learning algorithms and computing hardware nowadays are like human brains and neurons, while data is like human experiences. Human experiences affect a human success or not a lot. For example, even though some students are not much smarter than the others, they are more likely to succeed because they can get good education experiences, read good books, travel the world, and broaden their horizon. In this way, we believed it is time for researchers to pay more attention to how to collect more datasets.

7.3 Limitations

To our best knowledge, the two datasets we collected are the largest on-body sensor-based human activity datasets, while they are still much smaller than other domains' data, such as the ImageNet dataset [16]. Thus, we call on people to pay more attention to on-body sensor-based HAR data collection. We believed that SenseCollect will facilitate researchers to collect on-body sensor-based human activity datasets more efficiently and easily. We believe SenseCollect will inspire researchers to conduct more research on on-body sensor-based HAR data collection.

Furthermore, SenseCollect focus on locomotion tasks, which may not be able to complete in front of a computer. We recommended participants wear smartglasses for watching real-time signal feedback. Researchers are highly recommended to develop systems and transmit real-time visualized signals to smartglasses. Further, improving engagement by playing music needs further study in the future. For example, we will build game systems for participants to perform gestures and activities, such as walking and running, then investigate how the games affect the engagement. Although SenseCollect guideline is helpful to collect human activities (e.g., walking, sitting) and human gestures (e.g., writing, typing), SenseCollect is not suitable for collecting health data, such as heartbeats and emotions. These areas may bring additional challenges that SenseCollect is not considered. But some guidelines in SenseCollect, such as pay for performance, are worth to be considered. For example, pay for performance may be helpful to motivate participants to place the sensors well or keep the devices charged all the time. We will further study how to collect health data in the future.

8 RELATED WORKS

8.1 Sensor Data Collection in HAR

In contrast to computer vision and speech transcription, where one can download large datasets from the Internet, on-body sensor-based human activity data collection is often performed by conducting user studies [7, 40, 55]. Typically, the participants in a study are asked to perform activities while wearing a sensing platform. One way to conduct this is to record video data in addition to sensor data, which is subsequently used for manual data annotation. However, it is difficult to synchronize the sensor and video data streams [37]. Many research works developed open mobile systems for sensor data collection, such as activities [24], emotions [51], and social behavior [4, 49]. These systems either did not support data annotation [36], or required users to do self report annotation after corresponding events [20, 24, 51]. However, self report annotation data have two main problems: the user returning tainted data (e.g., data filled out late after some events instead of immediately), and the user returning incomplete data or having a low response rate. The most common way is asking users to perform required activities through notification/transcription. [9, 12–15, 30, 44, 50]. SenseCollect collects data by asking users to perform activities through notification, and investigate the challenges of this method. In this way, SenseCollect is able to get a larger, accurate dataset collected in a short time rather than current practice which often takes months and ends up with small datasets that are error-prone. Additionally, We made a data collection assisted system to provide real-time sensing feedback. Some research works focused on data collection

procedure, such as how to select well-suited participants [39] to collect sensor data, how to optimize incentive budget for participants to collect data [52], providing incentives for collecting privacy data [45]. However, SenseCollect selected random participants to collect activity data and ascertained what factors complicate the on-body sensor-based HAR data collection. In this way, SenseCollect explored an efficient way to facilitate data collection.

Other approaches have explored alternative data collection methods that do not directly involve human participants. Kang et al. render a 3D human model on computer graphics software and simulate human activities [28]. The sensor data is extracted from the simulated human motion, and subsequently used to train the recognition models. However, it is challenging to simulate and design complex human activities realistically. Therefore, such methods typically only explore simple gestures and locomotion activities. Rey et al. [41] proposed to collect virtual sensor data from online videos and demonstrated the effectiveness of the virtual sensor data for recognizing fitness activities. Their approach computes the 2D pose motion for a single person in the video with a fixed camera viewpoint. However, it is difficult to train for a target real sensor with the synced video and accelerometer recordings, which transfers the changes in joint locations from the 2D scene to the three-axis accelerometer norm. IMUTube [29] generated data from the full IMU (three-axis accelerometer, gyroscope, and simulated magnetometer). It performed 3D motion estimation from videos with multiple people and scenes in the wild using camera motion tracking. Although it did not require synced video or wearable recordings, it is almost impossible to generate data of fine-grained gestures, such as finger snaps and micro finger writing. In contrast, SenseCollect directly involved human participants and collected data from human body. Furthermore, SenseCollect analyzed the challenges and explores the efficient way to collect data from the real world.

8.2 Tackling Sparse Data Problem

The relatively small size of labeled on-body sensor-based human activity datasets makes models quickly overfitting and does not support the application of complex model architectures. To solve the problem of small datasets, data augmentation techniques have been applied to prevent overfitting, increase variability and improve generalizability in the datasets. They involve techniques that systematically transform the data during the training process in order to make classifiers more robust to noise and other variations [35]. They artificially inflate the training data by using data warping, or oversampling [43]. Data warping includes geometric transformations such as rotations, cropping, and adversarial training. For time series classification, the data warping techniques include window slicing, window warping, rotations, permutations and dynamic time warping [19, 31]. Over a single method, it can also combine several of these transformations to further improve the performance. Um et al. demonstrated that combining three basic ways (rotation, permutation, and time warping) yields better performance than using a single method [48]. In [38], construction equipment activity recognition is also improved by combining simple transformations. While the data augmentation techniques improve the classification performance, they, ultimately, produce perturbed training samples. Therefore, they cannot provide for the variety in human movements obtained by collecting data from a large number of subjects.

Recently, data generation using either oversampling or generative adversarial networks (GANs) [43] have also been successfully introduced to sensor-based human activity recognition [53]. The GAN based techniques perform augmentation by sampling from the datasets distribution. However, they require huge amounts of data to train, and may suffer from training instability and non-convergence. Furthermore, there is little prior work studying data augmentation by GANs for wearable sensor data and their actual suitability for sensor-based human activity recognition remains to be shown. This makes it difficult to readily apply these generative networks to generate more data.

Another approach to deal with small labeled datasets includes transfer learning. Here, a base classifier (typically a neural network) is first trained on a base dataset and task. Subsequently, the learned features are re-purposed,

or transferred, to a second target network to be trained on the target dataset and task. In particular, if the target dataset is significantly smaller compared to the base dataset, transfer learning enables training a large target network without overfitting [54], and typically results in improved performance. However, it relies on large datasets for knowledge transfer. As a result, [47] have noticed that the adoption of deep learning methods in human activity recognition has not yet translated to the noticeable accuracy gains seen in other domains.

We tackle the problem of collecting real-world on-body sensor-based human activity datasets. SenseCollect facilitates researchers, especially graduate students, to collect on-body sensor-based human activity and exploit the efficient way to collect qualified datasets.

9 CONCLUSION

In conclusion, SenseCollect explored how to collect on-body sensor-based human activity datasets efficiently by a three-phase study. First, we did a survey and interviewed students researchers in this area to understand the difficulties of the on-body sensor-based HAR data collection. Second, we designed and collected a dataset example to further investigate the potential factors affecting the data collection. Third, the deployment study showed that it was much easier to recruit students to manage and collect data from the participants. And providing a sensing feedback system to assist data collection, playing music/ videos, and imposing data quality and collection period requirements are significantly helpful to the data collection. It is also unnecessary to monitor the entire data collection procedure. With these findings, we implemented a system to assist data collection and collected two large datasets in a short time. We made the system and the datasets publicly available.

REFERENCES

[1] [n.d.]. Control variable. https://en.wikipedia.org/wiki/Control_variable.

[2] [n.d.]. Network socket. https://en.wikipedia.org/wiki/Network_socket.

[3] [n.d.]. Perfect pangrams. http://www.fun-with-words.com/pang_example.html.

[4] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing* 7, 6 (2011), 643–659.

[5] Marc Bachlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M Hausdorff, Nir Giladi, and Gerhard Troster. 2009. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine* 14, 2 (2009), 436–446.

[6] Jeroen C.M. Barte, Arne Nieuwenhuys, Sabine A.E. Geurts, and Michiel A.J. Kompier. 2019. Motivation counteracts fatigue-induced performance decrements in soccer passing performance. *Journal of Sports Sciences* 37, 10 (2019), 1189–1196. https://doi.org/10.1080/02640414.2018.1548919 arXiv:https://doi.org/10.1080/02640414.2018.1548919 PMID: 30472919.

[7] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033–2042.

[8] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005* (2013).

[9] Wenqiang Chen, Lin Chen, Yandao Huang, Xinyu Zhang, Lu Wang, Rukhsana Ruby, and Kaishun Wu. 2019. Taprint: Secure text input for commodity smart wristbands. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[10] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Shwetak Patel, and John Stankovic. 2021. ViFin: Harness Passive Vibration to Continuous Micro Finger Writing with a Commodity Smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[11] Wenqiang Chen, Lin Chen, Meiyi Ma, Farshid Salemi Parizi, Patel Shwetak, and John Stankovic. 2020. Continuous micro finger writing recognition with a commodity smartwatch: demo abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 603–604.

[12] Wenqiang Chen, Lin Chen, Kenneth Wan, and John Stankovic. 2020. A smartwatch product provides on-body tapping gestures recognition: demo abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 589–590.

[13] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2018. Vitype: A cost efficient on-body typing system through vibration. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[14] Wenqiang Chen, Maoning Guan, Yandao Huang, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2019. A Low Latency On-Body Typing System through Single Vibration Sensor. *IEEE Transactions on Mobile Computing* 19, 11 (2019), 2520–2532.
[15] Wenqiang Chen, Yanming Lian, Lu Wang, Rukhsana Ruby, Wen Hu, and Kaishun Wu. 2017. Virtual keyboard for wearable wristbands. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–2.
[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[18] Jialin Fan and Andrew P. Smith. 2017. The Impact of Workload and Fatigue on Performance. In *Human Mental Workload: Models and Applications*, Luca Longo and M. Chiara Leva (Eds.). Springer International Publishing, Cham, 90–105.
[19] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2018. Data augmentation using synthetic data for time series classification with deep residual networks. *arXiv preprint arXiv:1808.02455* (2018).
[20] Jon Froehlich, Mike Y Chen, Sunny Consolvo, Beverly Harrison, and James A Landay. 2007. MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services*. 57–70.
[21] Barry Gerhart and Meiyu Fang. 2015. Pay, intrinsic motivation, extrinsic motivation, performance, and creativity in the workplace: Revisiting long-held beliefs. (2015).
[22] Yuval Noah Harari and A Sapiens. 2014. A brief history of humankind. *Publish in agreement with The Deborah Harris Agency and the Grayhawk Agency* (2014).
[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[24] John Hicks, Nithya Ramanathan, Donnie Kim, Mohamad Monibi, Joshua Selsky, Mark Hansen, and Deborah Estrin. 2010. AndWellness: an open mobile system for activity and experience sampling. In *Wireless Health 2010*. 34–43.
[25] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29, 6 (2012), 82–97.
[26] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 493–504.
[27] Keith Jones. 2005. Music in factories: a twentieth-century technique for control of the productive self. *Social & Cultural Geography* 6, 5 (2005), 723–744.
[28] Cholmin Kang, Hyunwoo Jung, and Youngki Lee. 2019. Towards Machine Learning with Zero Real-World Data. In *The 5th ACM Workshop on Wearable Systems and Applications*. 41–46.
[29] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.
[30] Gierad Laput and Chris Harrison. 2019. Sensing fine-grained hand activity with smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[31] Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. 2016. Data augmentation for time series classification using convolutional neural networks. In *ECML/PKDD workshop on advanced analytics and learning on temporal data*.
[32] Daniyal Liaqat, Mohamed Abdalla, Pegah Abed-Esfahani, Moshe Gabel, Tatiana Son, Robert Wu, Andrea Gershon, Frank Rudzicz, and Eyal De Lara. 2019. WearBreathing: Real World Respiratory Rate Monitoring Using Smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–22.
[33] Jemma Looyestyn, Jocelyn Kernot, Kobie Boshoff, Jillian Ryan, Sarah Edney, and Carol Maher. 2017. Does gamification increase engagement with online programs? A systematic review. *PLOS ONE* 12, 3 (03 2017), 1–19. https://doi.org/10.1371/journal.pone.0173403
[34] I Scott MacKenzie and R William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In *CHI'03 extended abstracts on Human factors in computing systems*. 754–755.
[35] Akhil Mathur, Tianlin Zhang, Sourav Bhattacharya, Petar Velickovic, Leonid Joffe, Nicholas D Lane, Fahim Kawsar, and Pietro Lió. 2018. Using deep data augmentation training to address software and hardware heterogeneities in wearable and smartphone sensing devices. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 200–211.
[36] Md Abu Sayeed Mondol, Ifat A Emi, Sirat Samyoun, M Arif Imtiazur Rahman, and John A Stankovic. 2018. WaDa: An Android Smart Watch App for Sensor Data Collection. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 404–407.
[37] Thomas Plotz, Chen Chen, Nils Y Hammerla, and Gregory D Abowd. 2012. Automatic synchronization of wearable sensors and video-cameras for ground truth annotation–A practical approach. In *2012 16th international symposium on wearable computers*. IEEE,

100–103.
[38] Khandakar M Rashid and Joseph Louis. 2019. Times-series data augmentation and deep learning for construction equipment activity recognition. *Advanced Engineering Informatics* 42 (2019), 100944.
[39] Sasank Reddy, Deborah Estrin, and Mani Srivastava. 2010. Recruitment framework for participatory sensing data collections. In *International Conference on Pervasive Computing*. Springer, 138–155.
[40] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*. IEEE, 108–109.
[41] Vitor Fortes Rey, Peter Hevesi, Onorina Kovalenko, and Paul Lukowicz. 2019. Let there be IMU data: Generating training data for wearable, motion sensor based activity recognition from monocular rgb videos. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 699–708.
[42] Philipp M Scholl, Matthias Wille, and Kristof Van Laerhoven. 2015. Wearables in the wet lab: a laboratory system for capturing and guiding experiments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 589–599.
[43] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data* 6, 1 (2019), 1–48.
[44] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*. Springer, 510–526.
[45] Adish Singla and Andreas Krause. 2013. Incentives for privacy tradeoff in community sensing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 1.
[46] Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1029–1040.
[47] Catherine Tong, Shyam A Tailor, and Nicholas D Lane. 2020. Are Accelerometers for Activity Recognition a Dead-end?. In *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*. 39–44.
[48] Terry T Um, Franz MJ Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 216–220.
[49] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 3–14.
[50] Kaishun Wu, Yandao Huang, Wenqiang Chen, Lin Chen, Xinyu Zhang, Lu Wang, and Rukhsana Ruby. 2020. Power saving and secure text input for commodity smart watches. *IEEE Transactions on Mobile Computing* (2020).
[51] Haoyi Xiong, Yu Huang, Laura E Barnes, and Matthew S Gerber. 2016. Sensus: a cross-platform, general-purpose system for mobile crowdsensing in human-subject studies. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 415–426.
[52] Haoyi Xiong, Daqing Zhang, Guanling Chen, Leye Wang, and Vincent Gauthier. 2015. Crowdtasker: Maximizing coverage quality in piggyback crowdsensing under budget constraint. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 55–62.
[53] Shuochao Yao, Yiran Zhao, Huajie Shao, Chao Zhang, Aston Zhang, Shaohan Hu, Dongxin Liu, Shengzhong Liu, Lu Su, and Tarek Abdelzaher. 2018. Sensegan: Enabling deep learning for internet of things with a semi-supervised framework. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 3 (2018), 1–21.
[54] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792* (2014).
[55] Mi Zhang and Alexander A Sawchuk. 2012. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 1036–1043.