Published in final edited form as:
Proc Annu Symp Comput Hum Interact Play. 2019 October ; 2019: 135–147. doi: 10.1145/3311350.3347153.

Designing Videogames to Crowdsource Accelerometer Data Annotation for Activity Recognition Research

Aditya Ponnada, **Seth Cooper**, **Binod Thapa-Chhetry**, **Josh Aaron Miller**, **Dinesh John**, **Stephen Intille**
Northeastern University, Boston, MA

Abstract

Human activity recognition using wearable accelerometers can enable *in-situ* detection of physical activities to support novel human-computer interfaces and interventions. However, developing valid algorithms that use accelerometer data to detect everyday activities often requires large amounts of training datasets, precisely labeled with the start and end times of the activities of interest. Acquiring annotated data is challenging and time-consuming. Applied games, such as human computation games (HCGs) have been used to annotate images, sounds, and videos to support advances in machine learning using the collective effort of "non-expert game players." However, their potential to annotate accelerometer data has not been formally explored. In this paper, we present two proof-of-concept, web-based HCGs aimed at enabling game players to annotate accelerometer data. Using results from pilot studies with Amazon Mechanical Turk players, we discuss key challenges, opportunities, and, more generally, the potential of using applied videogames for annotating raw accelerometer data to support activity recognition research.

Keywords

Applied games; human computation; accelerometers; data annotation; activity recognition; crowdsourcing

INTRODUCTION

Sensor-based activity recognition algorithms use data from mobile and wearable devices to detect physical activities and other behaviors of people in everyday settings [9, 16, 49]. The data are collected from miniature accelerometers embedded in phones, smartwatches, fitness bands, and special-purpose activity trackers used for research studies. Recognition algorithms process raw data gathered from three-axis, high-sampling-rate wearable accelerometers using extracted signal features (e.g., frequency) and then detect/classify activities of interest to researchers [24, 49]. The output from these algorithms may include step counts [26], labels of specific activities [42], or energy expenditure [53]. Activities of special interest are physical activities, sedentary behaviors, and sleep (e.g., [50]). The field is highly active, with researchers proposing algorithms based on popular techniques such as support vector machines, decision trees, and ensemble methods, as well as newer variants on

ponnada.a@husky.neu.edu.

neural-network-based deep learning (e.g., [58]). In most cases, and especially with deep learning, these algorithms require large training datasets, where periods of raw data have been labeled with the target activities that were performed during data collection. Researchers typically gather data from devices worn by participants in controlled settings and then annotate the precise start and end times of each target activity (e.g., [8, 23, 56]). However, manually annotating raw accelerometer data is challenging and time-consuming [64]. As the need for large amounts of high-quality, labeled training data has intensified, researchers have started to explore new data annotation tools/methods to expedite activity recognition research [64].

Therefore, we are investigating the potential of using videogames to annotate accelerometer data to support research in activity recognition. Videogames offer engaging ways of solving complex problems. Videogame players have advanced science by solving puzzles (e.g., [17, 32, 39]) and by annotating images, videos, and other forms of data (e.g., [38, 41, 62]). In fact, crowdsourcing such tasks using applied human computation games (HCGs) allows a large number of casual players to collectively accomplish scientific tasks that are computationally intractable [61] by leveraging the skills of game players [6, 59]. Therefore, in this paper, we present two web-based HCGs; we use these proof-of-concept game prototypes to demonstrate how videogames played by non-expert players in the “crowd” can be used to annotate triaxial accelerometer data gathered from a wearable wrist sensor – the type of data increasingly used for activity recognition from wearable sensors. Players observe the data and using game mechanics, annotate the start and end times of the everyday physical activities (e.g., walking or sitting) performed at the time of data collection. We present the preliminary evaluations of our game prototypes using Amazon Mechanical Turk and discuss the game design challenges and potential of using videogames to crowdsource accelerometer data annotation. In addition, we explore the game design opportunities of gathering raw data annotations using the joint effort of activity recognition algorithms and labels from casual videogame players.

BACKGROUND

Our game design prototypes are based on prior research in accelerometry, annotation practices, and HCG designs.

Raw Accelerometer Data

Accelerometers (e.g., the commonly used ActiGraph in physical activity research) measure acceleration along the X, Y, and Z axes [1, 45], typically between ±2 and ±16 *g (g* = 9.8 ms-2). Accelerometer data can be used to infer the amount or type of physical activity a person wearing the sensor may have engaged in (e.g., [11-13, 43]). Accelerometers are popular in research studies because they are affordable, low-powered, and easy to maintain [45]. These devices can be comfortably worn on locations such as the wrist, ankle, hip, or thigh, and can be used to collect data for days or weeks. Different movements generate distinctive patterns. For instance, ambulation activities such as running or walking generate a rhythmic pattern with spikes from a wrist-sensor (e.g., Figure 1, top). Similarly, sedentary activities such as sitting or resting generate low-amplitude, non-rhythmic wrist movements (e.g., Figure 1, bottom). When patterns are sufficiently distinct, different activities can be automatically detected by an algorithm (or a trained expert). Higher sampling rates (e.g., 20-80 Hz) allow detection of movement changes with high temporal fidelity (e.g., every 10 s), whereas, lower sampling rates (e.g., 0.1 Hz) may be sufficient for detecting some types of prolonged activities (e.g., sleeping for hours).

Current Accelerometer Annotation Practices

In a typical training data collection session, participants are asked to perform a set of real-world activities and researchers log timestamps of these activities (e.g., [8]). Often, researchers use manual notes, spreadsheets, or custom-built software for annotation. Observers must be trained to annotate data, which require time and resources not available to every research group. Because of the effort and cost involved, most annotated datasets are small – involving fewer than 100 participants performing only a handful of activities for only a few minutes each (e.g., [8, 37]). Using cross-validation techniques, it is possible to train supervised learning algorithms on these small datasets, but machine learning algorithms work best with large, high-quality training data (e.g., [63]).

Thus, research is underway to reduce the effort required to annotate sensor datasets. For example, Barz et al. developed a multimodal multi-sensor annotation tool that uses video and raw sensor data to assist researchers annotating data retrospectively [10]. Similarly, Diete et al. developed a semi-supervised labeling tool for video and inertial sensor annotation, where a semi-supervised learning algorithm provides labeling assistance [25]. However, the tools designed for data annotations have only focused on gathering annotations from a single expert, which may introduce bias. Thus, the challenges of subjective biases, scalability of data, cost, and annotation consistency loom large for researchers, but are yet to be formally addressed [36, 64].

Applied Videogames for Data Annotation

Videogames can be used to leverage the problem-solving abilities of players to solve computationally challenging tasks [18, 20]; such applied games are known as HCGs [6, 59, 61]. They have been used to accomplish tasks such as labeling objects in images (e.g., ESP game [62] and KissKissBan [29]), annotating audio files (e.g., Tag-A-Tune [38]), and classifying animal species (e.g., Forgotten Island [46]). In these games, players use their common-sense knowledge, and the gameplay encourages player inputs that can be used to build knowledge important to researchers. Data are filtered for accuracy through techniques such as player agreement [47]. The labeled output from these games then could serve as training data for machine learning algorithms [61]. Other HCGs also teach new skills as part of the game, for example, protein and RNA folding (e.g. Foldit [17, 33] and EteRNA [39]), mapping the 3D structure of neurons (e.g. Eyewire [34]), or gene-disease annotation (e.g. Dizeez [41]). In these games, the players learn new skills through tutorials, enabling them to creatively solve challenging problems that are typically solved only by domain experts. Like the tasks targeted by these skill-training HCGs, understanding and labeling accelerometer data requires experience observing the raw signal for different activities before those signals can be differentiated correctly. The task is challenging because the accelerometer signal can include complexities such as ambiguous activity transitions (e.g., sitting to sleeping) and wear-location effects (e.g., acceleration measured from the wrist when cycling), which even experts can find difficult to identify and annotate. In fact, HCGs could be cost-effective compared to hiring domain-experts for similar tasks (e.g., ~500K Foldit players so far have played protein molecular puzzles for free).

In non-accelerometer data domains, several citizen science games have trained players to annotate complex data such as finding patterns in mammal sounds (e.g., Bat Detective [2] and Whale FM [5]), labeling transit photometry data of planets (e.g., Planet Hunters [4] and Project Discovery's Exoplanets [54]), and aligning gene sequences (e.g., Phylo [32] and Fraxinus [48]). Thus, it may be possible to train casual game players to annotate accelerometer data using applied videogames (i.e. HCGs).

GAME PROTOTYPE REQUIREMENTS

To focus our investigation, we gathered data annotation requirements from an exercise physiologist who processes accelerometer data using activity recognition algorithms, and a game design researcher (both co-authors of this paper). These discussions allowed us to iterate through preliminary game ideas and extract the basic requirements for accelerometer data annotation games intended to help players label data to be used in activity recognition research, especially for large wrist-worn datasets such as NHANES and UK Biobank [57]. These requirements relate to identifying the activities of interest for annotation, visualizing accelerometer data in the games, providing feedback to players using validation data, and gathering sample data to be used for pilot testing of the games.

Activities to Annotate Using the Applied Games

Physical activity researchers using raw accelerometer data are interested in classifying behavior into broad classes of physical activities, as well as specific activities (e.g., those listed in the Compendium of Physical Activities)[7]. Researchers are also interested in developing sleep detection algorithms using low-cost wrist-accelerometers (e.g., [28]). Moreover, detecting sensor non-wear for wrist-worn accelerometers can allow researchers to distinguish sensor wear data before training algorithms. Thus, for our initial game prototypes, we chose the following broad activities to annotate: *ambulation* (e.g., walking and running), *sedentary* (e.g., sitting, resting, working on PC), *sleep*, and *sensor non-wear*[35, 50]. However, the game designs ought to be sufficiently flexible to accommodate finer activity category labeling as well (e.g., brisk walking and bicycling).

Accelerometer Data Visualization for Games

Raw, triaxial (X, Y, and Z axes) accelerometer data visualization contains more information than summary data (e.g., step counts) computed from the raw data. Therefore, it may be necessary to train game players to annotate activity using raw accelerometer data [27]. Raw data visualization not only provides information on the intensity of movement during an activity, but also the orientation of the sensor. In fact, low-intensity activities such as sleep and sensor non-wear can be hard to distinguish from one another with summary data, whereas raw data may be used to visually differentiate between these activities.

Validation and Unknown Accelerometer Data for Games

When annotating raw accelerometer data, the game players must differentiate between different behaviors just by inspecting the raw data. Thus, instructions on using raw data representations of activities to annotate, followed by practice with corrective feedback on actual examples, could help players develop the required skills (e.g., as in the Foldit and EteRNA games). For this purpose, we employed the strategy of integrating data with known annotations with unlabeled data into the gameplay. A block of either type of accelerometer data will be referred to as a *fragment*. In our games, we will refer to the two different data-types:

- *Validation data* have pre-assigned labels (i.e., ground-truth) pre-programmed in the game to validate player responses. Validation data are also used in the tutorial phases of the game for training purposes.
- *Unknown data* are the accelerometer data to be labeled by the players.

This approach of simultaneously mixing validation and unknown data into gameplay is similar to the approach used by reCAPTCHA [60], where the validation and unknown text are displayed together to verify inputs on validation text and get human input on the unknown text in captchas.

Sample Accelerometer Data for Game Testing

Wrist-based accelerometers are increasingly being used in activity recognition research because this body location results in higher wear compliance and comfort as compared to ankle, thigh, or hip locations. In addition, the NHANES and UK Biobank studies have collected wrist-worn data from over 115K participants, collectively [15, 57]. Thus, to evaluate our games, we chose a raw accelerometer dataset from one of the authors' unpublished studies, where 50 participants wearing the accelerometer (Actigraph, 80 Hz) on the wrist carried out a protocol of real-world activities such as walking, sitting, and typing. Researchers annotated the activities in real time using a custom, tablet-based application. A subsample of these data was used as *validation data* in the games with the pre-programmed labels, and another subsample was stripped of annotations and used in the games as *unknown data*. Because the true labels for all the data are known, player performance in labeling the *unknown data* can be assessed post hoc.

GAME PROTOTYPES: OVERVIEW

We designed two games – *Mobots* and *Signaligner*. *Mobots* is an action-based game, where small moving fragments of data are labeled by shading/coloring one at a time (inspired by Guitar Hero [3]). *Signaligner* is a pattern-matching puzzle game, where players take their time to cut and align fragments of matching activities (inspired by Phylo [32]). The key design challenge in both the games is to enable players to correctly label the unknown data, and to provide feedback using validation data to improve their labeling skills. The purpose of designing two different games was to explore different genres (e.g., action vs. puzzle), mechanics (coloring vs. pattern alignment), and pace (fast vs. self-exploratory) for the accelerometer data annotation task.

Mobots: An Action Annotation Game

Mobots is similar to a rhythm-based game, where players annotate short windows of accelerometer data. Fragments of data (unknown and validation) travel along a track (from right to left) and the players must *press and hold* (the core game mechanic) a button on the keyboard (e.g., F for "walk/run") as the signal fragment crosses a line in order to color (i.e., annotate) the data fragment with a label (Figure 2). Players can change labels midway through a data fragment by pressing a different key if multiple labels are required for that fragment (e.g., transitions between activities). This game mechanic was chosen to enable independent labeling of small, zoomed-in fragments of data. Accuracy of labeling is assessed by examining player performance in labelling the seeded validation data. Players, however, are not informed about the presence of different types of data (validation and unknown data) to ensure unbiased labeling during the gameplay. Each fragment displayed in the game constitutes 10 s of data (sampled at 16 Hz). Different game levels ask players to identify different activities. Levels introduce new activities gradually so as to maximize player learning about how the data for different labels appear in the raw signal. The player's goal is to successfully complete as many levels as possible.

Game-progression—When a validation data fragment is labeled correctly, a green "power-bar" in the level is increased (Figure 2, bottom of left panel) and the pace of the moving fragments increases, making the level more challenging to play. The level is accomplished when the power bar fills up completely. The game provides data fragments until the level is completed or the player quits or pauses the level. Subsequently tougher levels are longer in duration with more data fragments (both validation and unknown) to annotate.

In-game feedback—When a fragment of data is not labeled, or when a fragment of validation data is labeled incorrectly, that fragment is highlighted and on-screen text feedback (e.g., "Please try again" or "This was Ambulation") highlights the correct annotation for that fragment (Figure 2, left). An incorrect label on validation data decreases the level's power bar, delaying level completion.

Game tutorials—The first three levels of the game are tutorial levels and use only validation data to teach the game mechanics that allow the player to label fragments. These levels teach: 1) how to label a fragment correctly, 2) how to change keys to label different activities in different fragments, and 3) how to change keys to label different activities within the same fragment. Following these levels, whenever a new activity label (e.g., *ambulation*) is introduced, it is done so using its own dedicated level comprising only of the validation data of that activity. This allows players to learn the visual pattern of this new activity before being presented with unknown data to label. At the beginning of each level, an instruction screen provides a description of what each category of label represents, complemented with a video of the activity and its acceleration pattern (Figure 2, right panel). This instruction screen is always accessible to the players during gameplay.

Signaligner: A Pattern Matching Puzzle Game

Signaligner is a pattern-matching puzzle game that allows players to view hours of data at once and determine the best annotations for different time windows. Players can use the mouse pointer to *cut, slide*, and *join* data fragments on the screen, labeling them with a background color; the core game mechanics are like the mechanics in typical sound editing software, such as Adobe Soundbooth. When cut, the fragment is divided into two, and then players can drag to slide the cut fragments across the screen. When a fragment is moved to abut another fragment, the fragments merge into one fragment again. This mechanic permits players to match acceleration patterns in multiple fragments of raw data that are stacked vertically. Some fragments can have validation data; this information is hidden from the players during gameplay, similar to Mobots. We chose a sampling rate of 0.2 Hz for the game's data to be able to display ~60 min of data at once on the web. Unlike Mobots, Signaligner does not train players to match fragments with activities directly; rather, it asks them to match the visual patterns with the template patterns provided for different activities (Figure 3).

The game checks if fragments are aligned vertically. The player can click on the background of this column to color it (with the background color of the sample patterns), indicating that all fragments in this column share the same activity label (and that the activity the data represent is different from data labeled with different colors).

Game progression—The goal in each level is to organize the cut fragments into independent columns, where each column (colored with one color) represents a particular activity label (Figure 3, left). At any time, players can click the "Check!" button at the bottom of the screen to verify if all the fragments are correctly organized. The game proceeds to the next level if all the validation fragments are correctly aligned. The harder levels contain more variety, where data from more activity types are mixed up. A player's overall goal is to successfully complete the final level.

In-game feedback—If the players have mislabeled a fragment that has validation data (e.g., if a column has two different activities' data), they receive feedback; the background color of the validation data fragment is highlighted, and correct label is displayed.

Game tutorials—Tutorial levels, which contain *only* validation data, include instructional text about how to complete the level. Players can re-check their solutions in tutorial levels as many times as they wish and then proceed to the next level once they have a correct solution. These levels were intended to both train the players and increase the likelihood that players who reach their challenge level could label data accurately. The game tutorial sequence introduces the game's mechanics as follows, with one tutorial level for each stage. Players must 1) change the assigned label by recoloring the fragment, 2) split a fragment into multiple fragments and reassign a new label, 3) un-align stacked fragments into different columns so they can be assigned different labels, and 4) split, un-align, as well as relabel multiple fragments (Figure 3, left panel). Following the tutorial levels, players are given one challenge level, randomly selected from a pool of challenge levels (e.g., Figure 3, right panel). Each challenge level has at least one validation and one unknown data fragment.

Players only have one chance to submit a solution to their final challenge level, after which they finish the game.

GAME PROTOTYPE EVALUATION

We assessed our game designs with players from Amazon Mechanical Turk (MTurk). MTurk is a commonly used platform to crowdsource tasks in human-subjects research (including the HCGs).The goal was to assess the feasibility of using videogames to gather annotations on raw accelerometer data.

Game Evaluation Methodology

We used MTurk and its TurkPrime interface [40] to recruit crowd players and the games were made accessible to players via a hyperlink. Player labels and interaction time were logged to a remote database with a unique identifier for each play session.; Each consenting player on MTurk was given a token code for $0.50 and was allowed to submit the code for payment *before or after* playing the game. We provided the code before the game to ensure that 1) there was no external motivation while playing and 2) players could play the game for as long as they wanted, with no pre-defined completion time [52]. In addition, Signaligner players also received an additional bonus of $0.50 for completing the challenge level. Finally, players also had the opportunity to provide optional, open-ended feedback about the game they played via a survey, but no additional compensation was provided for this feedback.

Game Evaluation Variables

The games were assessed for feasibility using labeling accuracy and inter-player labelling agreement. In addition, we captured play-session times and gathered subjective user experiences from the survey.

Play-session times—We measured session times as total time spent playing the game (including game tutorials) and time spent on the labeling tasks (excluding the tutorials).

Inter-player agreement—The inter-player agreement, or player consensus (between 0 and 1), was estimated as the proportion of the most frequently labeled activity from all the player labels for a given second of unknown data. For instance, if a given second of unknown data had four *ambulation* votes and one *sedentary* vote, then the final annotation was *ambulation* with an inter-player agreement of 0.8. In Mobots, unknown data samples with less than five player labels were not considered labeled. Higher inter-player agreement indicates more labeling consensus from independent players [47].

Labeling accuracy—Labeling accuracy measures correctness of the player annotation on the unknown data in the games compared with ground-truth labels. Labeling accuracy is assessed for the smallest time-window (1 s for Mobots and 5 s for Signaligner) aggregated (with consensus) from all the players for that sample [21, 47, 55]. Higher accuracy indicates better quality labeling of unknown data.

Subjective game experience—Subjective game experience was assessed qualitatively based on the open-ended feedback from the survey, which asked the players to describe their most positive and negative experiences of playing the games.

Exploration with an Activity Recognition Algorithm

We also compared labels from players with label output by an activity recognition algorithm to explore 1) how an algorithm predicts activities in the same unknown data used in the game and labeled by players, and 2) design opportunities for algorithm-assisted annotation games for activity recognition. Although physical activity researchers are now commonly collecting raw wrist-accelerometer data, as of writing this paper, there are few sufficiently validated algorithms that use raw wrist accelerometer data to predict labels of specific activities. We first modified the algorithm by Mannini et al. [42] that uses support vector machines to classify ambulation, sedentary, and sleep activities. The model was trained on a 25-participant subset of the sample data used for pilot testing our games and had 85% overall accuracy with leave-one-subject-out (LOSO) testing. We then modified this algorithm to use a random forest (RF) classifier with frequency domain features (i.e., dominant frequency, power of dominant frequency, and total power); using 30 s windows of data, it yielded 99% accuracy using the same LOSO test. The RF model provides the likelihood of each classification for the 30-s window, and the classification with the highest likelihood is selected. Due to its superior performance, we chose the RF classifier against which to compare player labels obtained from the games.

However, our focus was not to evaluate a particular algorithm. In practice, applied games for crowdsourcing annotation would be deployed to help with labeling data only when machines are unable to confidently label certain data fragments. Thus, for any labeling task, it is likely that investigators will run one or more algorithms, assess the confidence levels of the resulting labels, and then deploy games to fix or verify the uncertain labels. Thus, we are most interested in situations where the algorithm is uncertain, and the players are not, and how to design games to elicit high-quality labels from players.

GAME PROTOTYPE EVALUATION RESULTS

We evaluated player labels on unknown data from both the games and then compared these labels with the algorithm-inferred labels on the same unknown data.

Mobots Game Evaluation Results

The Mobots game prototype was designed with 30 levels, including the three tutorials. We received labels for *ambulation* and *sedentary* activities, and *sensor non-wear*.

Mobots play session times—For the MTurk task with 100 assignments, there were 82 sessions played. Thus, some players entered their code without playing the game. Mobots players spent a total of 520.80 min (8.70 h) playing the game. This time includes the short game introduction scene (that took a mean of 5 s per player to view). Players spent a median of 3.10 min (IQR = 4.70 min, range = 0.11-39.70 min) per player playing the game, resulting in a median pay rate of $9.70 per hour for labeling activities. The median time spent playing per level, for levels with both validation and unknown data (i.e., non-tutorial levels), was 0.97 min (IQR = 0.67 min, range = 0.002-18.90 min). Three players reached level 20, and 52 players played until level four or higher; level four is the first level with validation and unknown data, after tutorials. Players labeled 9.50 min of the unknown data (with 5 or more player labels per sample) displayed in the game.

Mobots inter-player agreement—The inter-player agreement for all the labels on unknown data was 0.73; 0.76 for *ambulation*, 0.68 for *sedentary*, and 0.89 for *sensor non-wear* activities (Figure 4, bottom).

Mobots labeling accuracy—Labels from Mobots are aggregated across players through player consensus and compared with ground truth labels for each second of unknown data labeled (Table 1). Fifty-three seconds of sedentary data (17.7%; out of 5 min of actual sedentary data labeled) were mislabeled as *ambulation*. Even though the game did not have the opportunity to present any sensor non-wear unknown data fragments, we received 6 s of *sedentary* data labeled as *sensor non-wear* as well. Figure 4 shows a snapshot of data labeled (aggregated) aligned with the raw accelerometer data (16 Hz) and the corresponding ground-truth labels. The accuracy of labeling unknown data when compared with their ground truth labels was 89.7%; 100% for ambulation and 78.1% for sedentary labels.

Subjective game experience—Fifteen out of 82 players provided voluntary feedback on Mobots. Ten players reported contributing to research as their motivation to play the game. Five players reported accurately detecting different activities within a fragment as their most positive experience. However, 11 players reported having difficulty mastering keyboard key-pressing when they identified changes in a pattern within a fragment; they expressed they needed more practice. Two participants expressed a desire for more positive reinforcement in the game for longer levels. Seven participants reported that the in-level increase in fragment speed made it harder for the them to keep track of different keyboard keys to press.

Signaligner Game Evaluation Results

Labels were logged when the players clicked “Check!” in the challenge level (Fig 3). The unknown data fragments each contained 20-59 min of data, with individual activity fragments of 9-40 min in length. The validation data in each challenge level was 8.25 min long. We received labels on *ambulation*, *sedentary*, and *sleep* activities.

Signaligner play session times—Although the MTurk task had 100 assignments, there were a total of 148 sessions logged. It is possible that some players tried exploring the game before deciding whether to submit their code; logs indicate nearly all sessions that did not have a code submission ended within the first two tutorials. For completeness, we present data on all 148 sessions for a total of 11.69 h of play time. Players spent a median of 3.11 min (IQR = 3.60 min, range = 0.10-55.20 min) playing the tutorial levels. Those who reached their challenge level spent a median of 0.6 min (IQR = 0.42 min, range = 0.25-0.95 min) playing their challenge level. Fifty-five players reached their challenge level. Of these, nine were assigned the sleep-only level, three the sedentary-only level, five the ambulation-only level, four the sleep/sedentary level, eight the sedentary/ambulation level, nine the ambulation/sleep level, and 17 the level containing all three activities. Players who completed the challenge level played for median 3.60 min and those who did not played for median 4.0 min, resulting in an estimated median pay rate of $ 16.70 and $7.50 per hour respectively.

Signaligner inter-player agreement—Signaligner players labeled the unknown data fragments with an agreement of 0.94 (Figure 5, bottom); 0.96 for *ambulation*, 0.99 for *sedentary*, and 0.88 for sleep *activities*.

Signaligner labeling accuracy—Signaligner's logs allowed us to compare players' labels on unknown data fragments in three ways (Table 2): 1) using data from *all* players, regardless of whether they labeled the validation fragment correctly or not, 2) using data only from the individual players who labeled the validation fragment correctly (i.e., *trusted* players), and 3) using the most often chosen label per sample aggregated across all the trusted players who labeled the validation data fragment correctly (i.e., *trusted* players' *consensus*) (Table 3). Data were displayed at a lower sampling rate than Mobots (0.2 Hz), so the label comparison unit was a 5 s window. Figure 5 shows aggregated player annotations with the raw data and their corresponding ground truth labels on a 20-min sample used in the game. The players were collectively confident in differentiating between *sleep, sedentary*, and *ambulation* fragments in the unknown data. Players had an overall accuracy of 90.7% (from all the players), 94.6% (from trusted players), and 99.5% (from the trusted players' consensus). Labeling accuracy on the unknown data was higher when considering trusted players' consensus.

Subjective game experience—For Signaligner, we received feedback from 10 players. Four players expressed finding matching patterns in data fragments a positive experience. However, six players reported having difficulty getting used to the labeling instructions in the early stages. For instance, one player mentioned having difficulty performing cut, slide, and join actions using the touchpad on his/her laptop. However, four players reported cutting data fragments accurately to be challenging and expected the game to be more lenient when evaluating players' inputs.

Player labels vs activity recognition algorithm

The overall labeling accuracy of the algorithm on the unknown data used in the games was 89.9%; 90.6% for ambulation, 89.8% for sedentary, and 94.5% for sleep activity classification. We summarize the player and algorithm labeling for the unknown data (compared with its ground truth for each second) from both the games in Tables 4 and 5. In case of Mobots, the algorithm misclassified 128 s of unknown data (i.e., 22.40% of total unknown data labeled), out of which the Mobots players provided correct labels for 110 s (85.90% of 128 s) of unknown data. The mean inter-player agreement among Mobots players for these correct labels (where the algorithm misclassified) was 0.60 (SD = 0.20). Likewise, in Signaligner, the algorithm misclassified 690 s of the unknown data (i.e., 5% of the total unknown data labeled), out of which the Signaligner players provided correct labels for 510 s (73.9% of 690 s) of the data. The mean inter-player agreement among Signaligner players for these correct labels (where the algorithm misclassified) was 0.92 (SD = 0.12). In fact, the instances where algorithm had moderate likelihood (<0.66) of correct prediction and players provided correct labels for that unknown data, Signaligner players had higher mean inter-player agreement (0.98, SD = 0.05) compared to Mobots (0.49, SD = 0.04).

DISCUSSION

In this paper, we presented two HCG design prototypes intended to motivate players to annotate raw accelerometer data—Mobots and Signaligner (of different game genres, mechanics, and pace). We assessed their feasibility using MTurk players (summarized in Table 6).

Signaligner players may have had higher labeling accuracy and inter-player agreement than Mobots players because, in Signaligner, players could use more signal context (and visual patterns) to match with an on-screen reference (Figure 3) of a similar acceleration pattern using their pattern recognition abilities. However, in Mobots, players were labeling short (~10 s) fragments of data using only the memories of signals and their corresponding activity categories (described in the activity tutorial). In Mobots, this learning and retaining of activity signal characteristics may be challenging for players. Moreover, it is possible that the additional bonus for completing the challenge level in Signaligner might have motivated the players to label more unknown data compared to Mobots players.

Mobots displayed 10 s fragments of the data (at 16 Hz) that allowed players to identify small bouts of ambulation and sedentary behavior. These bouts can get ignored in the zoomed-out view of Signaligner, where 0.2 Hz data are displayed so a 1 h view is shown. Nevertheless, Signaligner seems more suitable to label activities such as *sleep*, long *sedentary* behaviors, or *sensor non-wear*. Likewise, with such small data fragments, Mobots players might not be able to reliably differentiate *sleep* and *sensor non-wear*, because doing so may require observing longer windows (e.g., 30-60 min); anecdotally, our expert typically needs this context. Signaligner players could label more data because each fragment contains 20-59 min of data, more than the 10 s windows in Mobots.

Mobots and Signaligner players both reported having difficulty accurately identifying transitions between activities. Both the games had low tolerance for errors at transitions and thus provided feedback when there were errors on the validation data. In Mobots, despite using zoomed-in data that might have made it easier to perfectly mark a transition using the raw signal, the moving fragments (designed to increase engagement) made it harder for the players to master the skill of accurately changing keyboard keys at activity transitions. In Signaligner, despite having a much slower and self-directed pace, the zoomed-out view might have obscured the precise transitions between one activity and another and back (possible to see in Mobots) making it harder for Signaligner players to be able to cut the fragments at the right places. One approach to improve this game experience might be to allow more tolerance for transitions in the initial levels/stages of the game, and then to decrease the tolerance as the players improve their labeling skills. Alternatively, players promoted to be experts (after extended gameplay) might be sent data with the difficult transitions after it has been flagged from novice players due to lack of labeling consensus (inter-player agreement).

Mobots level design could be perceived as repetitive. Although new labels were introduced in the first ten levels, thereby improving player knowledge and skills and providing new game content, in later levels new skills were not required. Later levels only increased in difficulty (combining increases in speed and more complex combinations of activities to label); new skills/requirements were not introduced – thereby potentially not creating a desired cognitive flow experience [22]. Alternatively, Signaligner allowed players to explore tutorial levels for as long as they wished with unlimited attempts. Therefore, Signaligner players could master their skills at their own pace and try the challenge level only after they felt confident about their labeling skills. This self-direction may have helped sustain interest.

Comparing players' labels with an activity recognition algorithm presents several game design opportunities. Fortunately, the algorithm could reliably detect activities for the bulk of the unknown data. Algorithm pre-processing will be required to process datasets from large studies such as the UK Biobank (~100k participants). However, this pilot study does suggest that there will be instances where game players can help to verify or fix computer-generated labels. We observed three ways an activity recognition algorithm and labels by casual game players could be combined to produce well-labeled training data. First, there were instances where the algorithm misclassified an activity and player annotations had high inter-player agreement (e.g., Figure 6, left (1)). In such cases, future games could filter these instances, assign more players to annotate them, and then provide these examples as new training data to potentially improve the algorithm. Second, there were instances where the algorithm made correct predictions with high confidence but the players' annotations where wrong (e.g., Figure 6, left (2)). Such instances can be used in future games to provide game feedback to the players when labeling unknown data using algorithm's labels. Third, there are instances where the algorithm made low-confidence predictions, but the players' agreement was high (Figure 6, right (3)). In such cases, the player annotations can be used in future games to confirm algorithm output, which is how these types of crowd-based labeling systems would be used in practice. Algorithms would take preliminary labeling passes on huge datasets such as the NHANES and UK Biobank. Then, the crowd would label data where the algorithms are uncertain. This strategy requires, of course, that the algorithms be capable of outputting not only inferred labels, but likelihood scores for those labels on the unknown data. After players additionally label data, experts might then use the same tools for an additional labeling pass on only the data where non-experts (i.e. algorithms and players) do not agree. Engaging domain-experts could create an algorithms-players-experts loop for annotating raw data to improve activity recognition [14, 44]. The resulting massive datasets might then be used to further refine new algorithms using data-hungry methods such as deep learning.

In brief, our pilot testing of Mobots and Signaligner game prototypes presents an opportunity for game designers and researchers who might work together to further explore videogames to crowdsource accelerometer data annotation.

LIMITATIONS AND FUTURE WORK

Our pilot study and game prototypes have several limitations. First, Mobots players labeled only 9.5 min of data compared to Signaligner (3.8 h). Mobots was serving small amounts

(10 s) of data in each fragment and ultimately, we decided this game mechanic would not be feasible to label large quantities of data. As discussed above, an algorithm-assisted labeling game could filter the target data fragments that need labels from players because of incorrect or low likelihood automatic labeling.

Second, both Signaligner and Mobots games were designed with fixed zoom (i.e., visualization of data subsampled at a constant rate for each game) that did not allow players to explore data on their own at different zoom levels. A dynamic data zooming/panning interface, given that a single week-long dataset has ~2 GB of data, is complex to engineer because it requires extensive signal precomputation and caching to function effectively. We have since built such a tool, however, that can quickly fetch, subsample, and display the high-sampling-rate accelerometer data during the game sessions to allow rapid data visualization at any desired scale. This tool will allow us to develop a new type of game that might combine the best components of Mobots and Signaligner—the ability to zoom out and use large amounts of signal context to label some activities, and the ability to zoom in to mark activity transitions precisely.

Finally, our purpose in this pilot study was to assess the feasibility of our game prototypes to gather annotations from casual game players. Moving forward, our game prototypes could also further explore game engagement elements such as reward structures and game economy for long-term play as well as traditional social computing tools such as discussion forums, collaborations, and leaderboards that allow players to form an active citizen science gaming community (e.g., [51]). As we continue development on our games, we aim to give the players a stronger connection to the science behind the games [19, 30, 31].

ACKNOWLEDGMENTS

The research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (NIH) under award number UH2EB024407. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The work was also supported by NU-TECH AWS credits award from Northeastern University, Boston, MA. We also thank Rohan Jahagirdar and Clarence Hann for developmental support.

REFERENCES

[1]. Accelerometer Technologies, Specifications, and Limitations: A presentation by ActiGraph at ICAMPAM 2013 ActiGraph.

[2]. Zooniverse. 2013 The Bat Detective. (Accessed on 2018 12/05); Available from: https://www.batdetective.org/#!/home.

[3]. RedOctane and Harmonix. 2005 Guitar Hero. (Accessed on 2018 12/27); Available from: https://www.guitarhero.com/game.

[4]. Zooniverse. 2018 Planet Hunters. (Accessed on 2018 12/10); Available from: https://www.zooniverse.org/projects/nora-dot-eisner/planet-hunters-tess.

[5]. Zooniverse. 2011 Whale FM. (Accessed on 2018 12/19); Available from: https://whale.fm/.

[6]. von Ahn Luis and Dabbish Laura. 2008 Designing games with a purpose. Commun. ACM, 51(8): p. 58–67.

[7]. Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett DR Jr., Tudor-Locke C, Greer JL, Vezina J, Whitt-Glover MC, and Leon AS. 2011 2011 Compendium of physical activities: A second update of codes and MET values. Med Sci Sports Exerc, 43(8): p. 1575–81. [PubMed: 21681120]

[8]. Anguita Davide, Ghio Alessandro, Oneto Luca, Parra Xavier, and Reyes-Ortiz Jorge Luis. 2013 A public domain dataset for human activity recognition using smartphones. in ESANN.

[9]. Bao L and Intille SS. 2004 Activity recognition from user-annotated acceleration data, in Pervasive Computing, Ferscha A and Mattern F, Editors., Springer-Verlag: Berlin p. 1–17.

[10]. Barz Michael, Mohammad Mehdi Moniri Markus Weber, and Sonntag Daniel, Multimodal multisensor activity annotation tool, in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct 2016, ACM: Heidelberg, Germany p. 17–20.

[11]. Bassey EJ, Dallosso HM, Fentem PH, Irving JM, and Patrick JM. 1987 Validation of a simple mechanical accelerometer (pedometer) for the estimation of walking activity. Eur J Appl Physiol Occup Physiol, 56(3): p. 323–30. [PubMed: 3569241]

[12]. Bouten CV, Koekkoek KT, Verduin M, Kodde R, and Janssen JD. 1997 A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. IEEE Trans. on Bio-Medical Engineering, 44(3): p. 136–47.

[13]. Bouten CV, Westerterp KR, Verduin M, and Janssen JD. 1994 Assessment of energy expenditure for physical activity using a triaxial accelerometer. Med. Sci. Sports Exerc, 26(12): p. 1516–1523. [PubMed: 7869887]

[14]. Callaghan William, Goh Joslin, Mohareb Michael, Lim Andrew, and Law Edith. 2018 MechanicalHeart: A Human-Machine Framework for the Classification of Phonocardiograms. Proc. ACM Hum.-Comput. Interact, 2(CSCW): p. 1–17.

[15]. U.S. Department of Health and Human Services. 2018 NHANES - National Health and Nutrition Examination Survey Homepage. [Website] (Accessed on 2018 Nov 24); Available from: http://www.cdc.gov/nchs/nhanes.htm.

[16]. Choudhury T, Consolvo S, Harrison B, Hightower J, LaMarca A, LeGrand L, Rahimi A, and Rea A. 2008 The mobile sensing platform: An embedded activity recognition system. Pervasive Comp., 7(2): p. 32–41.

[17]. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popović Z, and Foldit players. 2010 Predicting protein structures with a multiplayer online game. Nature, 466(7307): p. 756–60. [PubMed: 20686574]

[18]. Cooper Seth, A Framework for Scientific Discovery through Video Games. 2014: Association for Computing Machinery and Morgan &; Claypool. 133.

[19]. Cooper Seth, Sterling Amy L. R., Kleffner Robert, Silversmith William M., and Siegel Justin B., Repurposing citizen science games as software tools for professional scientists, in Proceedings of the 13th International Conference on the Foundations of Digital Games 2018, ACM: Malm&ouml;, Sweden p. 1–6.

[20]. Cooper Seth, Treuille Adrien, Barbero Janos, Leaver-Fay Andrew, Tuite Kathleen, Khatib Firas, Snyder Alex Cho, Beenen Michael, Salesin David, Baker David, and Popović Zoran, The challenge of designing scientific discovery games, in Proceedings of the Fifth International Conference on the Foundations of Digital Games 2010, ACM: Monterey, California p. 40–47.

[21]. Costa J, Silva C, Antunes M, and Ribeiro B. 2011 On using crowdsourcing and active learning to improve classification performance. in 2011 11th International Conference on Intelligent Systems Design and Applications.

[22]. Csikszentmihalyi Mihaly, Finding flow: The psychology of engagement with everyday life. 1997: Basic Books.

[23]. De la Hoz E, Ariza P, Medina J, and Espinilla M, Sensor-based datasets for human activity recognition – A systematic review of literature. Vol. PP. 2018 1–1.

[24]. DeVaul RW and Dunn S, Real-Time Motion Classification for Wearable Computing Applications. 2001, MIT Media Laboratory.

[25]. Diete A, Sztyler T, and Stuckenschmidt H. 2017 A smart data annotation tool for multi-sensor activity recognition. in 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops).

[26]. Esliger DW, Probert A, Gorber SC, Bryan S, Laviolette M, and Tremblay MS. 2007 Validity of the Actical accelerometer step-count function. Med. Sci. Sports Exerc, 39(7): p. 1200–4. [PubMed: 17596790]

[27]. Freedson PS, Melanson E, and Sirard J. 1998 Calibration of the Computer Science and Applications (CSA), Inc. accelerometer. Med. Sci. Sports Exerc, 30(5): p. 777–781. [PubMed: 9588623]

[28]. Full Kelsie M., Kerr Jacqueline, Grandner Michael A., Malhotra Atul, Moran Kevin, Godoble Suneeta, Natarajan Loki, and Soler Xavier. 2018 Validation of a physical activity accelerometer device worn on the hip and wrist against polysomnography. Sleep Health: Journal Nat. Sleep Found, 4(2): p. 209–216.

[29]. Ho Chien-Ju, Chang Tao-Hsuan, Lee Jong-Chuan, Hsu Jane Yung-jen, and Chen Kuan-Ta. 2010 KissKissBan: a competitive human computation game for image annotation. SIGKDD Explor. Newsl, 12(1): p. 21–24.

[30]. Iacovides Ioanna, Jennett Charlene, Cornish-Trestrail Cassandra, and Cox Anna L. 2013 Do games attract or sustain engagement in citizen science?: a study of volunteer motivations. in CHI'13 Extended Abstracts on Human Factors in Computing Systems ACM.

[31]. Jennett Charlene, Kloetzer Laure, Schneider Daniel, Iacovides Ioanna, Cox Anna, Gold Margaret, Fuchs Brian, Eveleigh Alexandra, Methieu Kathleen, and Ajani Zoya. 2016 Motivations, learning and creativity in online citizen science. Journal of Sci. Comm, 15(3).

[32]. Kawrykow A, Roumanis G, Kam A, Kwak D, Leung C, Wu C, Zarour E, Sarmenta L, Blanchette M, and Waldispuhl J. 2012 Phylo: a citizen science approach for improving multiple sequence alignment. PLoS One, 7(3): p. e31362. [PubMed: 22412834]

[33]. Khatib F, DiMaio F, Cooper S, Kazmierczyk M, Gilski M, Krzywda S, Zabranska H, Pichova I, Thompson J, Popović Z, Jaskolski M, and Baker D. 2011 Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nat Struct Mol Biol, 18(10): p. 1175–7. [PubMed: 21926992]

[34]. Kim JS, Greene MJ, Zlateski A, Lee K, Richardson M, Turaga SC, Purcaro M, Balkam M, Robinson A, Behabadi BF, Campos M, Denk W, andSeung HS. 2014 Space-time wiring specificity supports direction selectivity in the retina. Nature, 509(7500): p. 331–6. [PubMed: 24805243]

[35]. Kosmadopoulos A, Darwent D, and Roach GD. 2016 Is it on? An algorithm for discerning wrist-accelerometer non-wear times from sleep/wake activity. Chronobiology International: p. 1–5.

[36]. Krüger F, Heine C, Bader S, Hein A, Teipel S, and Kirste T. 2017 On the applicability of clinical observation tools for human activity annotation. in 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops).

[37]. Lara Oscar D. and Labrador Miguel A.. 2013 A survey on human activity recognition using wearable sensors. IEEE Communications Surveys and Tutorials, 15(3): p. 1192–1209.

[38]. Law Edith LM, Ahn Luis Von, Dannenberg Roger B, and Crawford Mike. 2007 TagATune: A Game for Music and Sound Annotation. in ISMIR.

[39]. Lee J, Kladwang W, Lee M, Cantu D, Azizyan M, Kim H, Limpaecher A, Yoon S, Treuille A, Das R, and R. N. A. Participants Ete. 2014 RNA design rules from a massive open laboratory. Proc Natl Acad Sci U S A, 111(6): p. 2122–7. [PubMed: 24469816]

[40]. Litman L, Robinson J, and Abberbock T. 2017 TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behav Res Methods, 49(2): p. 433–442. [PubMed: 27071389]

[41]. Loguercio S, Good BM, and Su AI. 2013 Dizeez: an online game for human gene-disease annotation PLoS One, 8(8): p. e71171. [PubMed: 23951102]

[42]. Mannini A, Rosenberger M, Haskell WL, Sabatini AM, and Intille SS. 2017 Activity recognition in youth using single accelerometer placed at wrist or ankle. Med. Sci. Sports Exerc, 49(4): p. 801–812. [PubMed: 27820724]

[43]. Meijer GA, Westerterp KR, Verhoeven FM, Koper HB, and ten Hoor F. 1991 Methods to assess physical activity with special reference to motion sensors and accelerometers. IEEE Trans Biomed Eng, 38(3): p. 221–9. [PubMed: 2066134]

[44]. Méndez Ana Elisa Méndez, Cartwright Mark, and Bello Juan Pablo, Machine-crowd-expert model for increasing user engagement and annotation quality, in CHI’19 Extended Abstracts. 2019, ACM Press.

[45]. Miller Jeff, Accelerometer technologies, specfications, and limitations. Actigraph, LLC.

[46]. Prestopnik Nathan and Souid Dania, Forgotten island: a story-driven citizen science adventure, in CHI '13 Extended Abstracts on Human Factors in Computing Systems 2013, ACM: Paris, France p. 2643–2646.

[47]. Quinn Alexander J and Bederson Benjamin B. 2011 Human computation: a survey and taxonomy of a growing field. in Proceedings of the SIGCHI conference on human factors in computing systems ACM.

[48]. Rallapalli G, Players Fraxinus, Saunders DG, Yoshida K, Edwards A, Lugo CA, Collin S, Clavijo B, Corpas M, Swarbreck D, Clark M, Downie JA, Kamoun S, Cooper Team, and MacLean D. 2015 Lessons from Fraxinus, a crowd-sourced citizen science game in genomics. Elife, 4: p. e07460. [PubMed: 26219214]

[49]. Ravi Nishkam, Dandekar Nikhil, Mysore Preetham, and Littman Michael. 2005 Activity recognition from accelerometer data, in Proceedings of Innovative Applications of Artificial Intelligence, Jacobstein N and Porter B, Editors., AAAI Press: Menlo Park, CA p. 1541–1546.

[50]. Rosenberger ME, Buman MP, Haskell WL, McConnell MV, and Carstensen LL. 2016 24 hours of sleep, sedentary behavior, and physical activity with nine wearable devices. Med Sci Sports Exerc, 48(3): p. 457–65. [PubMed: 26484953]

[51]. Rowles Thomas A.. 2013 Power to the people: Does Eterna signal the arrival of a new wave of crowd-sourced projects? BMC Biochem, 14(1): p. 26. [PubMed: 24148199]

[52]. Sarkar Anurag and Cooper Seth, Comparing paid and volunteer recruitment in human computation games, in Proceedings of the 13th International Conference on the Foundations of Digital Games 2018, ACM: Malm&ouml;, Sweden p. 1–9.

[53]. Shcherbina Anna, Mattsson C. Mikael, Waggott Daryl, Salisbury Heidi, Christle Jeffrey W., Hastie Trevor, Wheeler Matthew T., and Ashley Euan A.. 2017 Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. Journal of Pers. Med, 7(2): p. 3.

[54]. Team Psycho Sisters, Exoplanets: The next phase of project discovery, in Dev-Blogs. 2017: Eve Online.

[55]. Song Jinhua, Wang Hao, Gao Yang, and An Bo. 2018 Active learning with confidence-based answers for crowdsourcing labeling tasks. Knowledge-Based Systems, 159: p. 244–258.

[56]. Stikic Maja and Schiele Bernt, Activity Recognition from Sparsely Labeled Data Using Multi-Instance Learning, in Location and Context Awareness. 2009, Springer, Berlin, Heidelberg p. 156–173.

[57]. Sudlow Cathie, Gallacher John, Allen Naomi, Beral Valerie, Burton Paul, Danesh John, Downey Paul, Elliott Paul, Green Jane, Landray Martin, Liu Bette, Matthews Paul, Ong Giok, Pell Jill, Silman Alan, Young Alan, Sprosen Tim, Peakman Tim, and Collins Rory. 2015 UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med, 12(3): p. e1001779. [PubMed: 25826379]

[58]. Tong Chuxuan, Zhang Jinglan, Chowdhury Alok, and Trost Stewart G., An Interactive Visualization Tool for Sensor-based Physical Activity Data Analysis, in Proceedings of the Australasian Computer Science Week Multiconference. 2019, ACM: Sydney, NSW, Australia p. 1–4.

[59]. von Ahn L. 2006 Games with a purpose. Computer, 39(6): p. 92–94.

[60]. von Ahn L, Maurer B, McMillen C, Abraham D, and Blum M. 2008 reCAPTCHA: human-based character recognition via Web security measures. Science, 321(5895): p. 1465–8. [PubMed: 18703711]

[61]. von Ahn Luis. 2008 Human computation. in Proceedings of the 2008 IEEE 24th International Conference on Data Engineering IEEE Computer Society.

[62]. von Ahn Luis and Dabbish Laura, Labeling images with a computer game, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2004, ACM: Vienna, Austria p. 319–326.

[63]. Wang Jindong, Chen Yiqiang, Hao Shuji, Peng Xiaohui, and Hu Lisha. 2019 Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters, 119: p. 3–11.

[64]. Yordanova Kristina, Paiement Adeline, Schröder Max, Tonkin Emma, Woznowski Przemyslaw, Olsson Carl Magnus, Rafferty Joseph, and Sztyler Timo. 2018 Challenges in annotation of user

data for ubiquitous systems: Results from the 1st ARDUOUS workshop. arXiv preprint arXiv: 1803.05843.

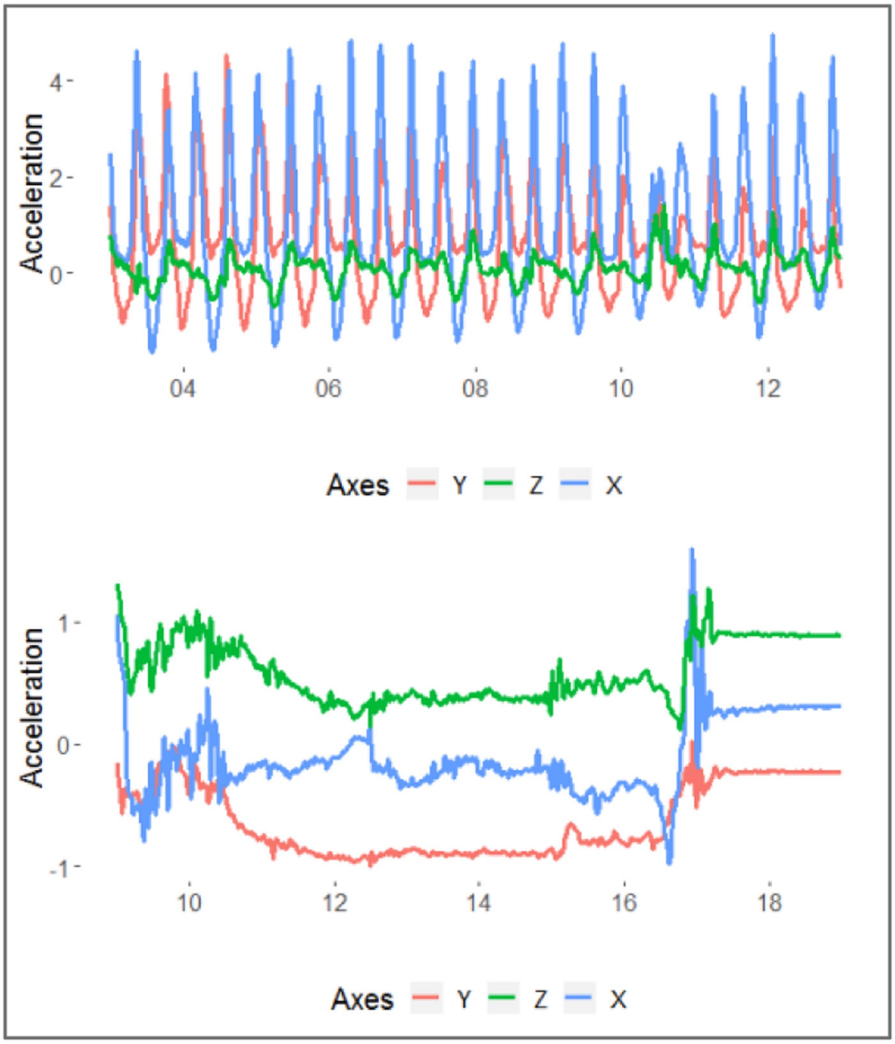
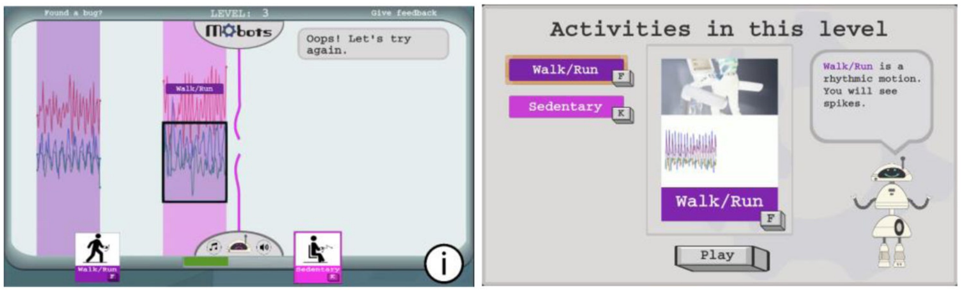


Figure 1.

10 s sample of raw acceleration (*g*) from the wrist for running (top) and sitting (bottom) activities.



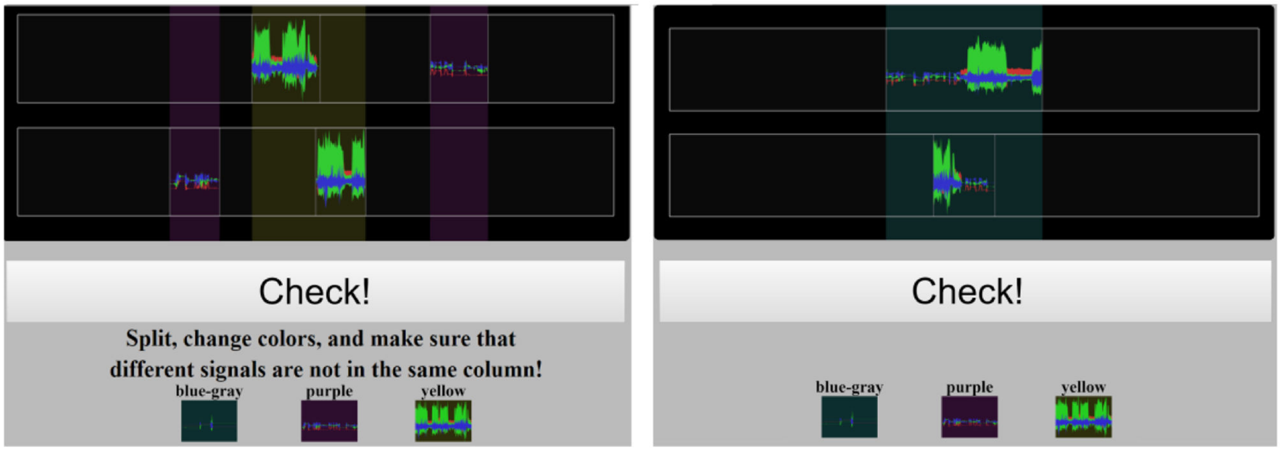
Oops! Let's try again.

Activities in this level

Walk/Run is a rhythmic motion. You will see spikes.

Figure 2.

Screenshots from Mobots. (Left) Highlights a player’s incorrect label on validation data fragment and shows the correct label. The green power bar at the bottom diminishes with player errors; (Right) Activity tutorial at the beginning of a level.



Split, change colors, and make sure that different signals are not in the same column!

Figure 3.

Screenshots from Signaligner. (Left) Completed state of the final tutorial level, labeled with correct activities in different columns; (Right) Starting state of the challenge level with one unknown data and one validation data fragment row. The sample signal-patterns to label are presented at the bottom with their unique background color for labeling.

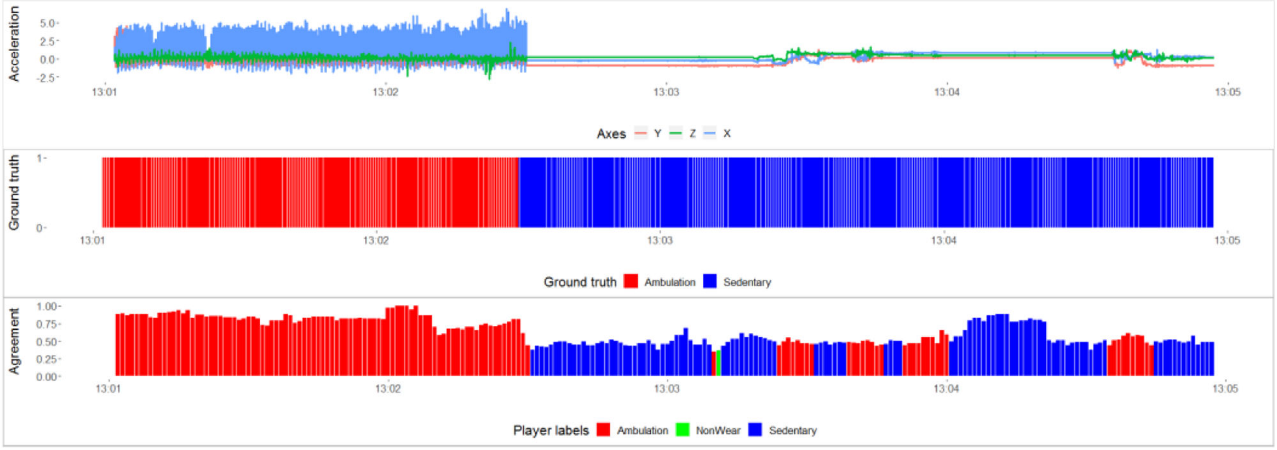


Figure 4.

A 4-min sample of raw data labeled by Mobots players; (top) Raw accelerometer fragment of unknown data; (middle) ground-truth labels of unknown data; (bottom) Player labels with inter-player agreement on the unknown data.

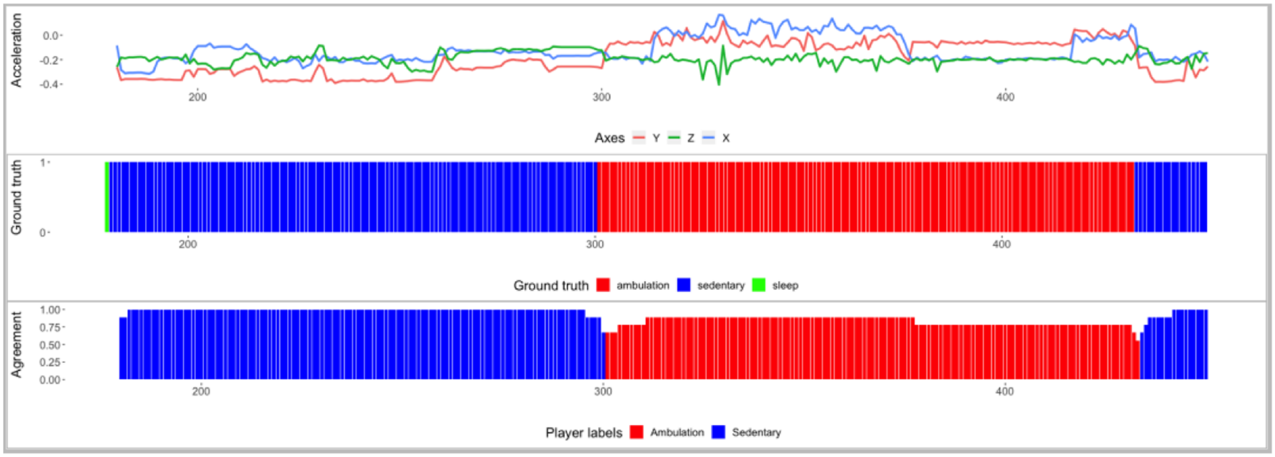


Figure 5.

20-min sample of raw data labeled by Signaligner players; (top) Raw accelerometer fragment of unknown data; (middle) ground-truth labels of unknown data; (bottom) Player labels with inter-player agreement on the unknown data.

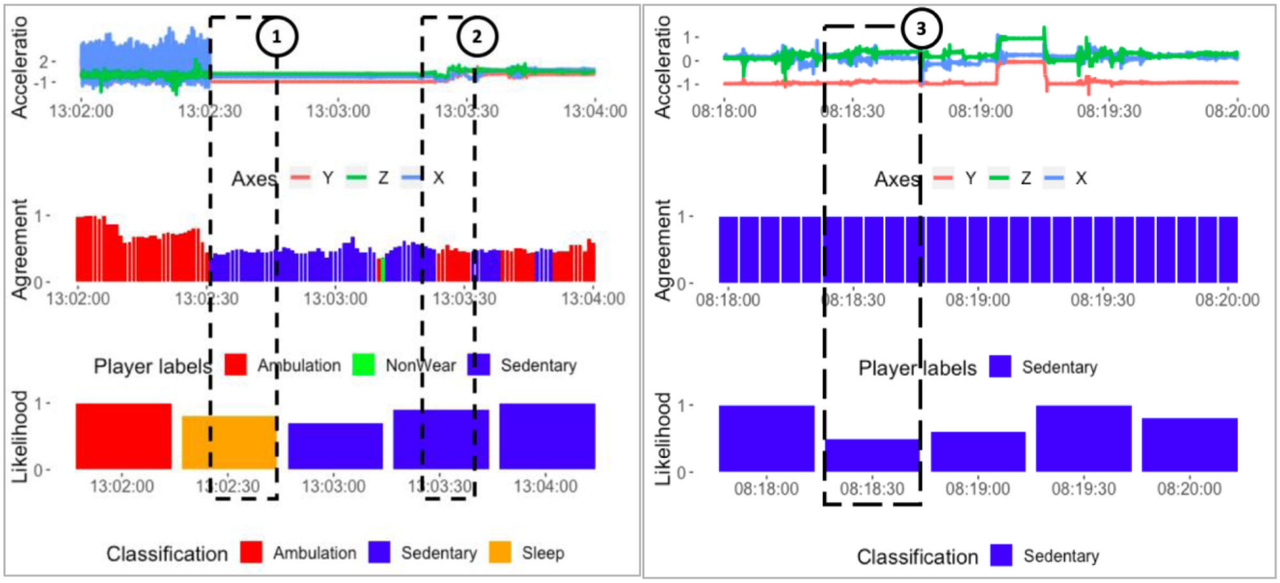


Figure 6.

Player labels on unknown data vs. activity recognition algorithm. (Left) 2-min sample used in Mobots game. In (1), the algorithm misclassified sedentary as sleep, but the players labeled it correctly. In (2), the algorithm was confident in classifying sedentary, but players incorrectly labeled it as ambulation. (Right) 2-min sample used in Signaligner game. In (3), the algorithm was less confident in classifying sedentary behavior, but players had high agreement for that label.

Table 1.

Comparing Mobots player annotations on unknown data with ground-truth labels for ambulation (amb.), sensor non-wear, and sedentary (sed.) activities.

Aggregate annotations		**Ground-truth reference**		
		Amb.	**Non-wear**	**Sed.**
Player annotations	Amb.	303	0	53
	Non-wear	0	0	6
	Sed.	0	0	211

Table 2.

Labeling accuracy on unknown data from all players, trusted players, and trusted players' consensus

	All players	Trusted players	Trusted players' consensus
Sleep	95.9%	88.3%	100%
Ambulation	89.7%	98.8%	98.8%
Sedentary	84.2%	98.8%	99.3%
Overall	90.7%	94.6%	99.5%

Table 3.

Comparing player labels on unknown data with ground truth reference from trusted players’ consensus

Trusted players’ consensus		**Ground truth reference**		
		Sleep	**Amb.**	**Sed.**
Player annotations	Sleep	1193	3	4
	Amb.	0	718	2
	Sed.	0	6	822

Table 4.

Players vs. algorithm labeling on the unknown data

Players vs. algorithm labeling for Mobots game		**Algorithm labels are**	
		Correct	**Incorrect**
Player labels are	Correct	401	110
	Incorrect	44	18

Table 5.

Players vs. algorithm labeling on the unknown data

Players vs. algorithm labeling for Signaligner game		**Algorithm labels are**	
		Correct	**Incorrect**
Player labels are	Correct	12525	510
	Incorrect	45	180

Table 6.

Comparison of Mobots and Signaligner games.

	Mobots game	**Signaligner game**
Game features		
Player goals	Annotate moving data fragments with activity names	Match fragment patterns with template visual patterns
Game type	Action	Puzzle
Displayed fragment size	10 s	20 – 59 min
Level completion	Fill the power bar through correct labels	Identify all activity validation fragments correctly
Raw data resolution	High (16 Hz)	Low (0.2 Hz)
Game labeling performance on the unknown data		
Data annotated	9.5 min	3.8 h
Play time	8.7 h	11.69 h
Inter-player agreement	0.73	0.94
Label accuracy	89.7%	99.5%
Sessions played*	82	148

*Testing with 100 target MTurk players