Capture24 Datasheet

1 Motivation

For what purpose was the dataset created? *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

This dataset was created to answer the need for a large-scale annotated dataset collected in the wild for accelerometer-based activity recognition tasks.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The lab group headed by Aiden Doherty at the Nuffield Department of Population Health, University of Oxford.

Who funded the creation of the dataset? *If there is an associated grant, please provide the name of the grantor and the grant name and number.*

Economic and Social Research Council, The British Heart Foundation Centre of Research Excellence, University of Oxford Advanced Research Computing, Oxford Biomedical Research Centre, Li Ka Shing Foundation, Health Data Research UK, Alan Turing Institute, National Institute for Health Research, Novo Nordisk, Medical Research Council.

Any other comments?
N/A

2 Dataset Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

The dataset collects time series of accelerometer data and corresponding activity labels describing the activities conducted by 151 participants in a 24-hour period. Demographic information (gender and age groups) is also provided for the 151 participants.

How many instances are there in total (of each type, if appropriate)?
There are 151 such time series. After applying a sliding window procedure to extract 10-second time windows, there are $n = 939,017$ windows.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

The data is a sample of the population of Oxfordshire, United Kingdom who volunteered to take part in the Capture-24 and Energy-24 Study in years 2014-2016. It cannot be regarded as a representative sample as the enrolled participants are on average younger and healthier than the overall population.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? *In either case, please provide a description.*

Raw acceleration in three axes with corresponding manually annotated text describing performed activities.

Is there a label or target associated with each instance? *If so, please provide a description.*

Yes, each instance has a fine-grained text annotation out of 206 possible annotations. We also provide six different schema for mapping the fine-grained annotations to coarse-grained annotations.

Is any information missing from individual instances? *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

The wearable camera was recorded at a lower sample rate (approx. 0.03Hz) than the accelerometer device (100Hz), which means the annotations were extrapolated to cover the entire time series. Also, images from the wearable camera were first reviewed by the participant to remove any sensitive images.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? *If so, please describe how these relationships are made explicit.*

N/A

Are there recommended data splits (e.g., training, development/validation, testing)? *If so, please provide a description of these splits, explaining the rationale behind them.*

Yes. We randomly selected 100 users as the training set. The rest are used as test.

Are there any errors, sources of noise, or redundancies in the dataset? *If so, please provide a description.*

Errors by human annotators cannot be ruled out. As mentioned above, the camera had a low sample rate, so there may be errors due to extrapolation. Also, wearable camera images can be hard to interpret at times.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

Yes, it is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? *If so, please provide a description.*

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? *If so, please describe why.*

No.

Does the dataset relate to people? *If not, you may skip the remaining questions in this section.*

Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

Only age groups (quartiles: 18-29, 30-37, 38-52, 53+) and sex (66% female) are provided as demographic information.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? *If so, please describe how.*

No. The data underwent extensive data processing to reduce the chances of reidentification.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? *If so, please provide a description.*

No.

Any other comments?
N/A

3 Collection Process

How was the data associated with each instance acquired? *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

Acceleration traces were directly recorded using wrist-worn activity trackers. Images captured by the wearable cameras were used to annotate the acceleration traces by trained human annotators. The cameras were worn during waketime only. For the sleep periods, this was based on self-reported sleep-wake times.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? *How were these mechanisms or procedures validated?*

Participants were asked to wear an Axivity AX3 wrist-worn triaxial accelerometer on their dominant hand. It was set to capture tri-axial acceleration data at 100 Hz with a dynamic range of ±8g. Wearable cameras were used to collect ground truths of the participants' activities while wearing the accelerometers. Participants were given a Vicon Autographer, a wearable camera which automatically takes photographs every 20 - 40 seconds, has up to 16 hours battery life and storage capacity for over one week's worth of images.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
It is not a subsample.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
Participants were recruited via advertisement in Oxford, United Kingdom. They were compensated for their participation with a £20 voucher.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? *If not, please describe the timeframe in which the data associated with the instances was created.*

Out of the accelerometer data from the total of 151 subjects, data from 132 subjects were collected as part of the Capture-24 Study in 2014-15, the rest were collected as part of the Energy-24 Study which took place in 2016.

Were any ethical review processes conducted (e.g., by an institutional review board)? *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

Yes. Capture24 received ethical approval from the University of Oxford Inter-Divisional Research Ethics Committee (Ref SSD/CUREC1A/13-262).

Does the dataset relate to people? *If not, you may skip the remaining questions in this section.*

Yes. Accelerometer and demographic data relating to 151 participants are included in this dataset.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
We collected the data from the individuals in question directly by handing out passive sensing devices (i.e. wrist-worn accelerometer and wearable camera) to them.

Were the individuals in question notified about the data collection? *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

Yes.

Did the individuals in question consent to the collection and use of their data? *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

Yes. A member of the research team fully explained the study requirements, after which participants signed a consent form. Consent procedures followed established ethical guidelines for wearable cameras in health behaviour research [2].

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

Yes.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

Yes.

Any other comments?

4 Data Preprocessing

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

The raw binary files from the activity trackers were processed to obtain CSV files. Gravity calibration and resampling were applied to the acceleration recordings to ensure uniformly good data quality across devices. The text annotations were revised and sensitive descriptions removed or simplified. For further anonymization, timestamps were randomized, and ages were converted to age groups.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? *If so, please provide a link or other access point to the "raw" data.*

All unprocessed data to reproduce the released dataset is securely stored and not publicly available. It may be available upon request.

Is the software used to preprocess/clean/label the instances available? *If so, please provide a link or other access point.*

https://github.com/activityMonitoring/biobankAccelerometerAnalysis

Any other comments?
N/A

5 Uses

Has the dataset been used for any tasks already? *If so, please provide a description.*

It has been used in a number of publications [1, 3, 4] resulting from epidemiological studies by members of our research group, but it has been made publicly available only recently.

Is there a repository that links to any or all papers or systems that use the dataset? *If so, please provide a link or other access point.*

https://github.com/activityMonitoring/capture24

What (other) tasks could the dataset be used for?
Aside from activity recognition, it could be used for energy expenditure prediction using the Metabolic Equivalent of Task (MET) values listed in the annotations (a regression problem). It could also be used to study open-set activity classification problems given the large number of different labels it contains.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

Ethics and privacy concerns were thoroughly considered. We do not believe there is a possibility of harm with the use of this dataset.

Are there tasks for which the dataset should not be used? *If so, please provide a description.*

Tasks seeking to re-identify the participants of the study.

Any other comments?
N/A

6 Dataset Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? *If so, please provide a description.*

The data has been made open-access in the website hosted by the Oxford University Research Archive.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) *Does the dataset have a digital object identifier (DOI)?*

Open-access website. DOI included.

When will the dataset be distributed?
The dataset has already been made publicly available.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

No.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? *If so, please describe these restrictions, and provide a link or other access point to, or*

otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

No.

Any other comments?
Note that the publicly available Capture24 dataset does not include the raw wearable camera images due to ethical concerns. Only the derived text annotations are provided.

7 Dataset Maintenance

Who will be supporting/hosting/maintaining the dataset?
The Oxford University Research Archive.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
E-mail address and telephone.

Is there an erratum? *If so, please provide a link or other access point.*

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

There are currently no plans for updating the dataset.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? *If so, please describe these limits and explain how they will be enforced.*

None.

Will older versions of the dataset continue to be supported/hosted/maintained? *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

There are currently no plans for hosting multiple versions of the dataset.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

Yes, they are welcomed to contact the creators of the dataset to discuss any extensions.

Any other comments?
N/A

References

[1] Aiden Doherty, Karl Smith-Byrne, Teresa Ferreira, Michael V Holmes, Chris Holmes, Sara L Pulit, and Cecilia M Lindgren. Gwas identifies 14 loci for device-measured physical activity and sleep duration. *Nature communications*, 9(1):1–8, 2018.

[2] P Kelly, S Marshall, H Badland, J Kerr, M Oliver, AR Doherty, and C Foster. Ethics of using wearable cameras devices in health behaviour research. *Am J Prev Med*, 44(3):314–319, 2013.

[3] Rosemary Walmsley, Shing Chan, Karl Smith-Byrne, Rema Ramakrishnan, Mark Woodward, Kazem Rahimi, Terence Dwyer, Derrick Bennett, and Aiden Doherty. Reallocating time

from machine-learned sleep, sedentary behaviour or light physical activity to moderate-to-vigorous physical activity is associated with lower cardiovascular disease risk. *medRxiv*, 2020.

[4] Matthew Willetts, Sven Hollowell, Louis Aslett, Chris Holmes, and Aiden Doherty. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 uk biobank participants. *Scientific reports*, 8(1):1–10, 2018.