DATASHEET:
Berkeley Open Extended Reality Recordings 2023 (BOXRR-23)

Vivek Nair
UC Berkeley

Wenbo Guo
UC Berkeley

Rui Wang
UC Berkeley

James O'Brien
UC Berkeley

Louis Rosenberg
Unanimous AI

Dawn Song
UC Berkeley

This document is based on *Datasheets for Datasets* [2].

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.

The dataset was originally collected for a research project on identifying users in VR [4]. The difficulty of uniquely identifying users is directly proportional to the number of users present; however, existing datasets of VR motion data only contained up to a few hundred users. To compare motion-based identification with traditional biometrics like facial recognition, a dataset of 50,000 or more users was required. Beyond user identification and authentication, we believe the data may be useful in a variety of fields, including motion synthesis, usability, and security/privacy.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset was created by the Center for Responsible, Decentralized Intelligence (RDI) at the University of California, Berkeley.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

This work was supported by the National Science Foundation, the National Physical Science Consortium, and the Fannie and John Hertz Foundation.

Any other comments?

Researchers interested in security and privacy applications of this dataset may be interested in our Data Privacy in Virtual Reality Systematization of Knowledge paper [1].

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.

Each instance consists of an XROR file that represents a recording of user motion data from an XR application.

How many instances are there in total (of each type, if appropriate)?

There are a total of 4,717,215 motion capture recordings (XROR files) included in the dataset.

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).

Yes – as of April 15th, 2023, the dataset includes all available recordings from the three sources we utilized.

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

Each XROR file includes a continuous sequence of frames recording the movement of a user's body parts in 3D space. It may also include metadata about the recording, as well as information about the events occurring in the XR application.

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

There is a variety of metadata associated with each XROR file, but no specific target label.

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.
All metadata originally associated with the recordings is retained, except for identifiable information such as user names and IDs, which were removed for privacy reasons.

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.
Yes – the recordings originating from the same user are always included in the same folder, with exactly one folder per user included in the dataset. Recordings originating from the same application, or the same in-application activity, are identifiable using the included metadata.

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.
For research involving user identification or authentication, we recommend splitting the data by recording, ensuring that one or more recordings for each user are present in each of the training, validation, and testing sets. Ideally, the data included in the training, validation, and testing sets should be as temporally distant as possible, as indicated by the timestamp metadata, to prevent session-specific attributes from being used for identification. For research involving user attribute inference, we instead recommend splitting the data by user, ensuring that no user is in more than one of the training, validation, and testing sets, so that models do not learn to identify users rather than inferring the attribute.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
All recordings present in the dataset are submitted directly by users around the world, using a wide variety of XR devices and technologies. As such, the data is highly heterogeneous and subject to a variety of sources of noise.

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.
No external data is required to utilize the data. Our libraries for XROR will remain publicly available via GitHub.

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 – all data was submitted by users with the knowledge and intention that it would be made publicly available.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Maybe – while there is nothing inherently offensive about the motion data, the metadata contains user-submitted text values and therefore may include offensive language.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes – the data records the motion of people while using XR devices and applications.

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.
Yes – the metadata identifies the country of each user, but other demographics like age and gender are not included.

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.
Maybe – original user pseudonyms and identifiers have been removed, but it may be possible to re-identify users via their motion data, such as by linking motion patterns to footage from surveillance cameras.

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.
Maybe – no sensitive data is directly included in the dataset, but it may be possible to infer sensitive attributes from user motion patterns.

Any other comments?
Attempting to link users in the dataset to their real-world identities, to contact these users, or to infer sensitive attributes from these users, is prohibited by the ethical data use agreement (DUA) accompanying this dataset.

COLLECTION

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.

For each recording, motion data, event data, and metadata were all automatically recorded by an application running on an XR device; there are no "guesses" in the data.

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. Finally, list when the dataset was first published.

The data was generated between November 1st, 2017 and April 15th, 2023 and was collected in April 2023.

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?

We used a number of Python scripts to download the data and metadata, and convert the files to the XROR format.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[6] for approaches in this area.)

The data was originally collected for purposes other than academic research. Assembling and processing the dataset required the use of a single workstation PC for about two weeks, with minimal associated cost and carbon footprint.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A – The data includes all available recordings from the three sources we utilized; no sampling was used.

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)?

Everyone involved in the data collection process is an author of this paper, and is either a student or professor at UC Berkeley. No additional compensation was provided.

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 – the collection and use of this dataset for academic research purposes has been reviewed and approved as protocol #2023-03-16120 by the Office for Protection of Human Subjects (OPHS) at UC Berkeley, an OHRP-certified Institutional Review Board (IRB).

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

Yes – the data records the motion of people while using XR devices and applications.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was originally submitted by users to three websites (BeatLeader, ScoreSaber, and Google Poly) with the knowledge that it would be made publicly available. We obtained and formatted the recordings directly from these three public sources.

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 – we made every attempt to notify the users of their involvement in academic research, including via official social media announcements from the original services.

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 – users explicitly consent to the publication of their recordings, and to their re-use for a variety of purposes, in the privacy policies and license agreements of the original services to which the recordings were submitted.

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 – a joint data removal procedure has been established for future iterations of the dataset, but it is not retroactive.

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 – an analysis of the potential impact of the dataset and its use on data subjects was conducted as part of the ethics review performed by the Office for Protection of Human Subjects (OPHS) at UC Berkeley.

Any other comments?
Regardless of the review conducted by OPHS, we encourage (and require) all research conducted using this dataset tp be independently reviewed and approved in advance by an IRB or equivalent ethics review board.

PREPROCESSING / CLEANING / LABELING

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.
Yes – minimal cleaning was performed to remove recordings that were corrupted or otherwise completely invalid.

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.
Yes – we have retained the raw data, but have not made it publicly available to protect the privacy of the users.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Yes – the XROR repository includes the code used to convert instances from the original BSOR, DAT, and TILT formats: `https://github.com/metaguard/xror`

Any other comments?
Overall, the data is preserved in as close to an original form as possible, while being compressed for efficiency and anonymized for ethical reasons.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
Yes – the data has been used to produce a VR identification study [4], a VR privacy study [5], and a VR privacy tool [3].

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.
Known uses of the dataset are identified on our website: `https://rdi.berkeley.edu/metaverse/boxrr-23/`

What (other) tasks could the dataset be used for?
The data could be used for a variety of purposes including human motion synthesis, cheating detection, score prediction, security & privacy research, human-computer interaction research, and machine learning research.

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?
While data was voluntarily submitted by users for broad public availability, knowledge of the implications of this data is currently limited. Thus, we advise caution when using the dataset for purposes that users may not have reasonably foreseen and could find objectionable.

Are there tasks for which the dataset should not be used? If so, please provide a description.
We prohibit attempting to link users in the dataset to their real-world identities, to contact these users, or to infer sensitive attributes from these users, as users may not understand the possibility of doing so.

Any other comments?
It is possible, and indeed likely, that there are uses for this motion capture data beyond what the researchers have imagined. We simply advise discretion with respect to the potential uses of data that originates from human subjects.

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.
Yes – the dataset can be accessed by any researcher who agrees to the license agreement and data use agreement.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset can be accessed via the following DOI: `https://doi.org/10.25350/B5NP4V`

When will the dataset be distributed?
The dataset is publicly available as of June 1st, 2023.

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.
Yes – researchers using this dataset must agree to our CC BY-NC-SA 4.0 license and ethical data use agreement.

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 – our redistribution of this data is permissible under the original licenses and applicable regulations, and our license is at least as restrictive as the original licenses.

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 – to our knowledge, there are no regulations limiting the international export of this data.

Any other comments?
Our eligibility to legally distribute this dataset has been confirmed by the Intellectual Property & Industry Research Alliances (IPIRA) office at UC Berkeley.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset is supported, hosted, and maintained by the Berkeley Center for Responsible, Decentralized Intelligence (RDI) as a public service at no cost.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
You can contact Berkeley RDI at rdi@berkeley.edu, or the primary author, Vivek Nair, at vcn@berkeley.edu.

Is there an erratum? If so, please provide a link or other access point.
No, but future errata, if applicable, will be included on our website: `https://rdi.berkeley.edu/metaverse/boxrr-23/`

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)?
Yes – new motion data is constantly being submitted, and we intend to update the dataset approximately once per year.

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.
No – users explicitly authorized the indefinite retention of all submitted recordings.

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.
Yes – we intend to keep all versions of the dataset publicly available for as long as possible.

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 – our license permits researchers to augment the dataset, as long as it is distributed under identical terms.

Any other comments?
Any updates to the responses contained in this document will be communicated via the official dataset page: `https://rdi.berkeley.edu/metaverse/boxrr-23/`

REFERENCES

[1] Gonzalo Munilla Garrido, Vivek Nair, and Dawn Song. Sok: Data privacy in virtual reality, 2023.

[2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.

[3] Vivek Nair, Gonzalo Munilla Garrido, and Dawn Song. Going incognito in the metaverse, 2022.

[4] Vivek Nair, Wenbo Guo, Justus Mattern, Rui Wang, James F. O'Brien, Louis Rosenberg, and Dawn Song. Unique identification of 50,000+ virtual reality users from head & hand motion data, 2023.

[5] Vivek Nair, Christian Rack, Wenbo Guo, Rui Wang, Shuixian Li, Brandon Huang, Atticus Cull, James F. O'Brien, Louis Rosenberg, and Dawn Song. Inferring private personal attributes of virtual reality users from head and hand motion data, 2023.

[6] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv:1906.02243 [cs]*, June 2019.