Datasheets for RVD: A Handheld Device-Based Fundus Video Dataset for Retinal Vessel Segmentation

Paper ID: 223

June 2023

1 Data Document

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.

Existing datasets for retinal vessel segmentation are image-based datasets collected with bench-top devices. The static images naturally lose the dynamic characteristics of retina fluctuation, resulting in diminished dataset richness, and the usage of bench-top devices further restricts dataset scalability due to its limited accessibility. Here, we provide a collection of 635 smartphone-based videos with detailed vessel annotation. The sequential frames significantly facilitate the analysis of subtle fluctuations in the retinal structure. The smartphone-based devices offer better flexibility and portability, allowing for scalable data collection. The diverse annotations provided meet the requirements for different ocular diseases diagnosis. Our RVD dataset is the first handheld device-based fundus video dataset for retinal vessel segmentation.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
This dataset is curated by the University of Queensland (UQ) computer vision group and the University of Technology Sydney (UTS) Eye-Vision team from four clinics, including Prince of Wales Hospital, Centre for Eye Health, Sydney Eye Hospital, and Broken Hill Hospital.

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.
The project is funded by Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC).

Any other comments?

None.

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.

Our RVD dataset is a video dataset with diverse annotations. Each video is associated with binary vessel masks, general artery-vein masks, fine-grained artery-vein masks, labels for SVP recognition and localization, and "peak" and "trough" annotations.

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

We collect 635 videos with 1,270 annotations for retinal vessel segmentation and 635 video annotations for spontaneous venous pulsation (SVP) detection and localization.

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).

Our dataset is the largest dataset for retinal vessel segmentation to date. It covers retinal videos from various patients and clinics. It is a newly curated dataset and does not involve instances from any existing datasets.

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

Original videos.

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

Yes. In our RVD dataset, we invite specialized experts to annotate each video with binary vessel masks, general artery-vein masks, fine-grained artery-vein masks, labels for SVP recognition and localization, and "peak" and "trough" 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.

No.

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. Each video has a unique ID. The videos from the same or different people are also distinguishable.

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 retinal vessel segmentation as well as temporal detection and localization

of SVP, we provide the splits of *training*, *validation*, and *testing* on our website.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
No.

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.
The dataset is self-contained. It does not link to external resources, *e.g.*, other datasets.

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, all videos are recordings of the fundus of human eyes.

Does the dataset identify any sub-populations (e.g., by age, gender)? If so, please describe how these sub-populations are identified and provide a description of their respective distributions within the dataset.
No.

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.

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?
None.

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

3

(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.

We collect all the retinal videos from four different clinics. Once the videos are obtained, we invite six specialized clinicians to annotate the binary vessel masks, general artery-vein masks, fine-grained artery-vein masks, labels for SVP recognition and localization, and "peak" and "trough" annotations.

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 design handheld devices by connecting a smartphone to the fundus camera lens. The clinics adopt the designed devices for data acquisition.

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

We curate our dataset from scratch. It is not from an existing dataset.

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)?

We invite the clinicians from four clinics to collect the videos. Then, we also ask them to annotate the collected 635 videos. The clinicians are also the authors of this work so they annotate the data for free.

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.

Our data collection starts from June, 2018. The whole collection process lasts five years.

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, we perform this study in accordance with the guidelines of the Tenets of Helsinki. Written consent was obtained from all participants prior to any data collection, and all examination protocols adhered to the tenets of the Declaration of Helsinki.

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

Yes, all videos are the recordings of fundus of human eyes.

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

We collect all the videos from four clinics. Clinicians are trained to operate the designed hand-held devices to examine patients' retinas while collecting fundus videos.

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. The participants are fully aware of data collection when they undergo their annual medical examinations.

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.

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, this is guaranteed by the guidelines of the Tenets of Helsinki.

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.

No.

Any other comments?
None.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/ labeling of the data done (e.g., dis- cretization 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. In our dataset, we utilize specific data cleaning methods to clean and preprocess the original video (see Sec. 3.2). The elimination of such noise improves the quality of our dataset and facilitates annotations.

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 provide the "raw" data on our website.

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

We use faster RCNN and the template matching algorithm for data cleaning and preprocessing. We use ImageJ and Adobe Illustrator for video annotation.

Any other comments?
None.

Uses

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

No, the dataset is newly proposed by our team.

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.

Yes, we provide the link to all related information on our website.

What (other) tasks could the dataset be used for?

The annotations of our dataset are diverse. Our dataset is not exclusively used for retinal vessel segmentation but also caters to an array of other retinal vessel-related tasks, *e.g.*, SVP recognition and localization.

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?

Our RVD dataset is able to improve the analysis of video data collected with handheld devices. The deployment of such portable devices dramatically extends the feasibility of eye examination or monitoring across diverse scenarios. This, in turn, promotes greater flexibility in data acquisition, facilitating the effortless collection of primitive videos and leading to a dataset with high extensibility. The augmented video dataset will further increase the diversity of the data, thereby boosting retinal vessel segmentation research in the future.

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

No.

Any other comments?

None.

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.

No.

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 will be publicly accessed on our website.

When will the dataset be distributed?

The dataset will be released to the public after acceptance. We provide a private link during the review process.

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.

We release our dataset under BY-NC-ND license.

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?

None.

Maintenance

Who will be supporting/hosting/ maintaining the dataset?

The University of Queensland (UQ) computer vision team.

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

E-mail addresses of the corresponding authors will be provided.

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)?

Yes. If errors are encountered, future versions of the dataset will be released and updated on our website.

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.

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.

Not applicable.

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.

We welcome the researchers to provide additional annotations. To ensure the quality and reliability of these new annotations, we will perform a rigorous validation process via a cross-validation mechanism. For instance, we will leverage the newly provided data as training data and our existing annotation as testing. Once the additional annotations have been verified, we will incorporate the new annotations into our dataset while acknowledging the contributors.

Any other comments?

No.