Datasheet of the LEG-3D-US dataset

Vanessa Gonzalez Duque[1,2,3,4*], Alexandra Marquardt[1,*], Yordanka Velikova[1,2], Lilian Lacourpaille[5], Antoine Nordez[5], Marion Crouzier[5], Hong Joo Lee[1], Diana Mateus[4], Nassir Navab[1,2,3]

[1] Computer − Aided Medical Procedure and Augmented Reality,
[2] Technical University of Munich, Garching bei Muenchen, Germany,
[3] Munich Data Science Institute, Munich, Germany,
[4] LS2N laboratory at Ecole Centrale Nantes, Nantes, France,
[5] MIP laboratory, EA4334, F − 44000 Nantes, France,

January 2024

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 LEG-3D-US dataset was recorded in 2018 during the study of Crouzier et al.[1] to investigate factors influencing muscle coordination during isometric tasks, focusing on triceps surae muscles. In 2020, the dataset was pre-proceeded and improved in the work of Duque et al. [2] for training deep-learning segmentation networks.

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

Data was created in close collaboration between the MIP laboratory of Nantes University and the LS2N laboratory of Ecole Central Nantes.

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.

This study was supported by two grants:

- The grant of QUETE project no. 2015-09035 from Pays de la Loire, a scholarship from the French ministry of higher education and research and fellowship from the Institut Universitaire de France (IUF).
- The European Regional Development. Fund, the Pays de la Loire region on the Connect Talent scheme (MILCOM Project) and Nantes Metropole (Convention 2017-10470).

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 instances are pairs of 3 dimensionals arrays (Witdh-Heigh-lenght) in MetaImage Medical Format (MHA) format containing the ultrasound compounded volumes and the 3D muscles labeled in integers [100,150,200].

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

There are 44 pairs of ultrasound volumes of the leg and 44 volume labels, filling voxel grids of 564×632×1443±(49×38×207), with an average isotropic voxel spacing of 0.276993 mm3 ± 0.015 mm3. Left and right legs were split 20 and 24 respectively.

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

** These authors contributed equally to this work.

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 dataset contains all the 3 possible instances of the muscles in 3D after post-processing of 2D available sparse annotations. With available annotations in any 2D slice of the volume.

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 pair of volumes contains a 3D ultrasound volume and a 3D label volume, created with Imfusion software [3] functions. The ultrasound volume uses a Voxel-Based Method reconstruction algorithm which assigns to the voxels the Nearest Neighbor pixel in the 2D images and fills the empty voxels using Gaussian weighting of 5 voxels around. Labels were obtained after a much-complicated process explained in the post-processing section.

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

Four labels are presented in a 3D label array with integer values $\{0, 100, 150, 200\}$ representing 0 the background, 100 the Soleus muscle (SOL), 150 the Gastrocnemius Medialis (GM), and 200 the Gastrocnemius lateralis (GL).

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.

Everything is included. No data is missing.

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.

There is no relationship between individual instances because each participant's labels correspond to its specific volume.

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.

We recommend keeping the split provided in the zip files for testing and validation because for 15 participants only borders were refined by an expert with an interpolation method called "ZOI" from the paper [2].

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

The training dataset generated using the "ZOI" method has approximately a 10% error rate, primarily at the muscle borders. In contrast, the test dataset, which underwent thorough verification and enhancement by an expert, has an error rate of less than 4%, similar to the error performed by experts in annotations over Computer tomography volumes.

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.

Dataset is self-contained. A copy of the dataset is saved at Ecole Centrale Nantes and at Nantes University.

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.

Unknown to the authors of the datasheet.

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, to the best of our knowledge.

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

Yes, the dataset relates to volunteers.

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.

Participants were aged between 18 and 45 years old, with height 173 ± 11 cm, and body mass of 64.3 ± 12.4

kg. The study contains data from 22 females and 22 males. Participants had no history of lower leg pain that had limited function or required them to seek intervention from a healthcare professional within the past 6 months previous to the study.

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.

To the best of our knowledge, it is not possible to identify individuals. Data is anonymized following the rules on health data in the light of the GDPR of the EU Member States[4].

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.

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.

2D muscles annotations were drawn over the 2D B-mode ultrasound images of size 221x537±(18×32) pixels with pixel spacing of 0.1723x0.1727±(0.037×0.045)cm using the Stradwin software[5] that allows to visualize the 2D images in a 3D context. Such 2D images are not available in this version of the dataset. Two double-blinded experts annotated the muscles twice to evaluate the intra-operative volumetric error, obtaining just 4% of the error. They had 1 year and 3 months of experience. Just the annotations of the more expert clinicians were used to create the 3D muscle models using the "ZOI" method.

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?

Ultrasound images were recorded at different frequencies with a 40mm linear VERMON probe: 2-10 MHz; and Aixplorer, Supersonic Imagine Ultrasound machine. Tracking of the ultrasound probe was performed with 6 cameras, using the Optitrack system-Natural point. Participants were prone with one leg in a custom-made bath to prevent pressure dependency in the measure. Four to six parallel sweeps were performed from the knee to the ankle. Images were recorded every 5 mm in low-speed mode. High-resolution 3D US volumes are compounded using the tracking matrices of the probe.

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

For 15 participants a second recording with different frequencies was acquired to visualize more details of the deeper muscle, the Soleus. In this version of the dataset, only the recordings with the higher frequency were made open source, providing one ultrasound per participant.

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)?

Dataset acquisition was led by Marion Crouzier under the supervision of Lilian Lacourpaille and Antoine Nordez. Data post-processing was performed by Vanessa Gonzalez under the supervision of Diana Mateus and Nassir Navab. Set up of the Webpage and maintenance is done by Alexandra Marquardt, Yordanka Velikova, and Hong Joo Lee.

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.

Ultrasound data collection spanned a 2-month period, while annotating 2D images using Stradwin software extended for 3 months. Achieving expert-level accuracy in creating 3D muscle models required an additional 4 months. For a total of 9 months of work.

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.

The experimental procedures were approved by the local ethics committee (Rennes Ouest V, CPP-MIP-010), and all procedures adhered to the Declaration of Helsinki.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
Data collection was done by the Motion Interaction and Performance group (MIP) at Nantes University. Data post-processing was done by the Laboratory of Numeric Sciences (LS2N) at Ecole Centrale Nantes.

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.
Participants were informed of the methods used before providing written consent. More details can be found at [1].

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.
Participants signed a written consent of their data for research use.

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).
Individuals provide their consent with no revoke mechanism.

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.

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.
The 3D labeling process is explained in more detail in the publication [2]. In summary, the "ZOI" method comprises a series of straightforward yet effective steps, ensuring high reproducibility while significantly reducing the manual annotation burden. It takes as input a collection of partial 2D annotations on high-resolution B-mode images, which serve as seed points. Employing a zero-order label propagation method, we copy the smallest mask to all unannotated images in between paired partial masks, followed by a Gaussian smoothing filter application for refinement. When two ultrasound volumes with different frequencies are available, we perform monomodal image-based rigid registration to align the two acquisitions, facilitating the transfer of annotations to enhance mask quality. The final correction of the borders is performed by an expert in 10% of the time taken to produce such annotation from scratch.

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.
The raw composed of tracking sweeps of 2D ultrasound images was saved but it will not be released in this version.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Ultrasound recording and 2D annotations were performed using the Stradwin software[5] and the Volume conpounding and 3D muscle creation was done using the Imfusion GmbH Suite[3].

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
One work has used the dataset already for developing deep learning architectures for automatic muscle segmentation. Refer to Alchanti et al-2021 [6].

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, refer to the GitHub repositorys:
https://github.com/DawoodChanti/IFSS-Net and https://github.com/Al3xand1a/segmentation-border-analysis

What (other) tasks could the dataset be used for?
This dataset can be used for many deep-learning tasks, for example, for network label variability evaluation or to create and analyze the way we evaluate segmentation performance as we did in Duque et al. 2024 IPCAI publication. Data can also serve to pre-train networks that will be fine-tuned in other ultrasound datasets. Additionally, the size of this dataset, when treated as 2D images slicing the 3D volumes, makes it interesting for labeling propagation.

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?
Volume compounding uses a Gaussian Kernel of 5 to fill the empty voxels decreasing the speckle patterns but increasing the quality. Sometimes speckle is used as a feature for experts. Please keep this in mind.

Are there tasks for which the dataset should not be used? If so, please provide a description.
The number of participants does not represent a valid sample from a population for clinical use, any created method needs to be verified in new participants— to make consequential decisions about people.

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 be distributed (e.g., tarball on the website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
Dataset can be downloaded on the webpage of Chair for Computer Aided Medical Procedures & Augmented Reality on the link: https://www.cs.cit.tum.de/camp/publications/leg-3d-us-dataset/

When will the dataset be distributed?
Dataset is made open-source on the 19th of January 2024.

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 free the data with a GNU General Public License (GPL). It requires that any derivative works also be open-source and distributed under the same terms. It ensures that any modifications to it remain open-source.

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.
All parties agree on the GPL license.

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.
N/A.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The dataset will be hosted by the Technische Universitat Munchen and maintained by the chair of Aided Medical Procedure and Augmented Reality(CAMP).

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Please contact us by email at vanessag.duque@tum.de or vgonzalezd@unal.edu.co

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)?
No, labeling error will not be updated, however, new instances from contributors could be checked and added if needed. Please contact us by email.

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.
N/A

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.
Maintenance will be done on the hosting webpage to make sure data can be downloaded correctly, but there will not be additional work on the labels. Any improvement of the annotations should be done by the users.

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.
The data in the training set contain a volumetric error of 10% in the training set, borders do not match with the images all the time due to the interpolation method used for creating the 3D models from the 2D sparse annotations. If others want to contribute, the refining of such annotations is welcomed.

References

[1] M. Crouzier, L. Lacourpaille, A. Nordez, K. Tucker, and F. Hug, "Neuromechanical coupling within the human triceps surae and its consequence on individual force-sharing strategies," *Journal of Experimental Biology*, vol. 221, no. 21, 2018.

[2] V. G. Duque, D. Alchanti, M. Crouzier, A. Nordez, L. Lacourpaille, and D. Mateus, "Low-limb muscles segmentation in 3d freehand ultrasound using non-learning methods and label transfer," in *16th International Symposium on Medical Information Processing and Analysis*, vol. 11583. SPIE, 2020, pp. 154–163.

[3] O. Zettinig, M. Salehi, R. Prevost, and W. Wein, "Recent advances in point-of-care ultrasound using the imfusion suite imfusion suite for real-time image analysis," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation: International Workshops, POCUS 2018, BIVPCS 2018, CuRIOUS 2018, and CPM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16–20, 2018, Proceedings*. Springer, 2018, pp. 47–55.

[4] J. Hansen, P. Wilson, E. Verhoeven, M. Kroneman, M. Kirwan, R. Verheij, and E.-B. van Veen, "Assessment of the eu member states' rules on health data in the light of gdpr," 2021.

[5] A. Gee, R. Prager, G. Treece, C. Cash, and L. Berman, "Processing and visualizing three-dimensional ultrasound data," *The British journal of radiology*, vol. 77, no. suppl_2, pp. S186–S193, 2004.

[6] D. Al Chanti, V. G. Duque, M. Crouzier, A. Nordez, L. Lacourpaille, and D. Mateus, "Ifss-net: Interactive few-shot siamese network for faster muscle segmentation and propagation in volumetric ultrasound," *IEEE transactions on medical imaging*, vol. 40, no. 10, pp. 2615–2628, 2021.