DMP title

Project Name Hyperspectral Imaging for Online Quality Control using Deep Learning (FWO DMP) - DMP title

Project Identifier 209870

Grant Title 1S28522N

Principal Investigator / Researcher Remi Van Belleghem

Description The goal of this project is to explore how deep learning networks can be used to perform quality control on biological product by using hyperspectral images. For this purpose, multiple datasets of hyperspectral images will be collected from plants, fruits and other biological materials. To analyze these datasets, deep learning networks with specific adaptations for hyperspectral images will be constructed and tested.

Institution KU Leuven

1. General Information

Name applicant

Remi Van Belleghem

FWO Project Number & Title

research project 1S28522N: "Online quality control of biological products and processes using hyperspectral imaging and deep learning"

Affiliation

- KU Leuven

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Generate new data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Type of Data	Format	Volume	How created
Raw hyperspectral linescan images	.tiff	1-4TB	Hyperspectral linescan camera (Specim FX10 and FX17). Measurements include dark and white reference
Processed hyperspectral images	.h5 or .hdr	1TB	White and dark corrected, cropped and filtered hyperspectral images based on R*aw hyperspectral linescan images* (see above)
RGB areascan images	.tiff, .png	50GB	High resolution RGB camera
Illustration and documentation images and videos	.jpg, .mp4	5GB	Cellphone images and videos acquired during experiment to serve for documenting experimental setup.
Deep learning network architectures	.py	1GB	Python files defining the created networks, written for Pytorch.

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- No

Privacy Registry Reference:

Short description of the kind of personal data that will be used:

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- No

If valorisation would become a possibility during this research, this valoration would involve the code that has been produced during this reserach, and not the data (hyperspectral images) acquired.

Do existing 3rd party agreements restrict dissemination or exploitation of the data

you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

4. Documentation and metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

1. For the RAW hyperspectral linescan images, camera settings used during datacollection are saved as an .icd file.

2. For the reconstructed hyperspectral images, a detailed README.md file will be included with the dataset. Furthermore, metadata like wavelengths corresponding to the hypercube channels, will also be stored in the image data itself using the hdf5 or hdr format. The dataset will also include .xml files that indicate which images where used for training, validation and testing of the algorithms.

3. The high resolution areascan RGB data will be linked to there corresponding hyperspectral image such that they can serve as a reference.

4. The python code will be extensivly documented in the code itself, to facilitate reading and understanding this code. Furthermore, a README file and minimal working example will be provided with the code.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- No

Metadata for the reconstructed hyperspectral images is already provided in the hyperspectral image data format (.h5 or envi). All other information, belonging to a entire dataset, will be included in the dataset's README file.

5. Data storage and backup during the FWO project

Where will the data be stored?

1. The RAW measurement data are stored on the measurement PC, and a backup is made to a hard -disk drive (HDD).

2. The reconstructed hyperspectral images, created form the RAW measurements, are stored locally on the processing PC, and a backup is made on a HDD. Furthermore, the scripts used for creating these reconstructed hyperspectral images from the RAW measurements are stored only using GitLab. At the end of the project, these clean, reconstructed hyperspectral images will be stored on the network storage from the reserach group (K:\SET-MeBioS-D0755\Biophotonics-0002).

3. GitLab repositories (https://gitlab.kuleuven.be) will be used as a backup for all code (.py and .ipynb files) generated. Output (graphs, images, ...) from the code will not be stored on the GitLab repositorie, but can be created using the code and RAW measurements. At the end of the project, these repositories will be handed over to the project supervisor.

4. All working documents (reports, powerpoints) are saved on OneDrive as backup.

How is backup of the data provided?

Because of the datasize of the RAW hyperspectral linescan images, hard-disk drives (4TB) will be used as a back up for the measurement data.

All code is backed up using GitLab remote repositories.

All working documents (reports, presentations) are backup in OneDrive, with an additional backup on the university's servers using SyncBack.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available then explain how this will be taken care of.

- Yes

Yes, multiple hard-disk drives from 4TB are available.

What are the expected costs for data storage and back up during the project? How

will these costs be covered?

Two 4TB hard-disk drives were already available in the lab, but additional drives can be purchased when necessary (100EUR-200EUR).

The price for storing data on the university's network drives, the price is 100EUR/TB/year.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

All Gitlab repositories are possible to configure in Private or Public mode, in orde to control who has acces to the code. The reason some code should be protected can be because of valorization potential.

The data that are stored on the network drive from the lab, can only be accessed by people within the lab. Because no personal data will be used, this protection is sufficient.

6. Data preservation after the FWO project

Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

All RAW data will be stored together with code needed for processing the data. Also the papers, presentations and other results belonging to a certain experiment will be kept with the data.

Where will the data be archived (= stored for the longer term)?

1. The data will be stored on the university's central servers (with automatic back-up procedures) for at least 10 years, conform the KU Leuven RDM policy.

2. For very large volumes (RAW data), hard disk drives will be used.

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

1. The storage of the research output can be done on the Large Volume Storage (LVS) from KULeuven (100EUR/TB/Year) at a cost of 1000EUR (2TB, 5Years). Because 2TB of data is available, processed hypercubes can also be stored on the LVS server.

2. RAW measurement data can not be stored on the LVS because size would be a problem, and therefore this data is stored on 2 hard disk drives (8TB) which together cost 200-400EUR.

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

At the moment, no restricting factors can be identified.

Which data will be made available after the end of the project?

1. Processed hyperspectral images that were used for training the algorithms will be made available. Because the dataset will be larger then 50GB, it is not trivial to host the data on Zenodo and a different open-acces cloud storage facility is required.

2. Code will be made available by the use of the Gitlab repositories.

Where/how will the data be made available for reuse?

- In an Open Access repository

1. The processed hyperspectral images will be uploaded in the ENVI format to an open-acces cloud storage server.
2. The source code will be released on Gitlab.

When will the data be made available?

Upon publication of a paper, the supporting dataset and source code will be made availble. One exception is when volarization potentional can be found in the dataset or source code.

Who will be able to access the data and under what conditions?

As all readers should be able to reconstruct the research results, both dataset and code will be made avaible to the under an open-acces framework. No restrication are given to acces or reuse of the code, as will be stated in an MIT license.

What are the expected costs for data sharing? How will the costs be covered?

Sharing of code can be done free of charge on Gitlab.

The hyperspectral images are to large to share using Zenodo (max 50GB) or RDR (50GB). Therefore, a different, non-free, solution will be sought. For example it is possible to use Google Cloud Storage at 0.02USD/GB/Month. An other option would be to share the data from Microsoft OneDrive (max 1TB).

8. Responsibilities

Who will be responsible for data documentation & metadata?

The doctoral student (Remi Van Belleghem) will be responsible to document all research that has been performed, and provide metadata for the data collected.

Who will be responsible for data storage & back up during the project?

Remi Van Belleghem will be responsible for providing backup of his data, code and research output during the project. For this he can make use of the tools provided within the research group.

Who will be responsible for ensuring data preservation and reuse ?

After the project, prof. Wouter Saeys will be the contact point for other researchers that want to acces the data and code that has been produced within this project. Therefore he will be responsible for ensuring preservation and reuse.

Who bears the end responsibility for updating & implementing this DMP?

The PI (prof. Wouter Saeys) bears the end responsibility of updating & implementing this DMP.