PID2021-125051OB-I00 HVD (2022-2025)

Harvesting Visual Data: enabling computer vision in unfavourable data scenarios

D3.1v1

Data Management Plan

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

AUTHORS LIST

José M. Martínez	josem.martinez@uam.es
Pablo Carballeira López	Pablo.carballeira@uam.es
Marcos Escudero Viñolo	Marcos.escudero@uam.es

HISTORY

Version	Date	Editor	Description
0.1	16/08/2023	José M. Martínez	First Draft version for contributions
0.2	16/08/2023	Marcos Escudero Viñolo	Contributions
0.3	06/09/2023	Pablo Carballeira López	Contributions
0.4	11/09/2023	Marcos Escudero Viñolo	Contributions
0.5	13/09/2023	Pablo Carballeira López	Final Working Draft editing
1.0	15/09/2023	José M. Martínez	Editorial checking

CONTENTS:

1. Introduction

This Document describes the Data Management Plan for the HVD project. We have created a data management template that will be completed for each dataset that is generated within this project. The data management plan for each dataset will collect specific information about its characteristics and common data management policies that apply to all datasets.

The document contains currently the following chapters:

- Chapter 1: Introduction. This short introduction.
- Chapter 2: Data Management Plan Template.
- Chapter 3: Data Management: Dataset (tbc)

Future versions of this document will include Chapter 3 and more, one for each HVD created dataset.

2. Data Management Plan Template

The data management plan is structured into eight sections. Some of them describe data storage, access, and preservation measures that apply to all the datasets created along the project. The other sections are specific for each dataset and describe data, annotation, privacy, and ethical characteristics for each dataset. See disambiguation between common and specific sections in the section title.

2.1. Data Description (specific)

This section provides a detailed description of the dataset, including the type of image or video sequences, the resolution, the number of images, the duration of each sequence, and the specific objects/information you have annotated.

2.2. ata Collection (specific)

This section explains how you have collected the data. Provide information about the equipment you have used, such as cameras or sensors. Include details about the data capture process, sampling frequency, and any calibration steps involved.

2.3. Annotation Process (specific)

This section describes how the annotation process has been. Specify the annotation methodology, whether it involves manual annotation by human annotators or automated techniques. If human annotators are involved, discuss their training process and the quality control measures you will implement.

2.4. Data Storage (common)

This section provides details on how the collected datasets are stored. It considers the amount of storage required and specific hardware infrastructure used. Additionally, it addresses the backup strategies and data redundancy measures implemented to ensure data integrity. As data storage is a critical component of the data management plan, a common strategy has been designed for all datasets. The following measures will be implemented to ensure secure and reliable storage of the project's data:

2.4.1. Professional Data Storage Servers

Dedicated professional data storage servers will be utilized to store the project's data. These servers have the necessary capacity, performance, and scalability to accommodate the anticipated volume of data. Redundancy measures will be implemented to ensure data availability and protection against hardware failures or disasters. RAID (Redundant Array of Independent Disks) technology will be utilized to distribute data across multiple disks, providing fault tolerance and increased data reliability. Regular data backups will be performed to create additional copies of the data for recovery purposes. Hash algorithms or checksums will be used to detect any data corruption or unauthorized modifications. Periodic integrity checks will be performed to verify the consistency and accuracy of the stored data.

2.4.2. Physical Access Restrictions

The servers will be housed in a secure and controlled environment: the server-room of the Escuela Politécnica Superior of the Universidad Autónoma de Madrid. Physical access to the data storage servers will be strictly limited to authorized system administrators only.

2.4.3. Storage Management

An efficient and organized storage management system will be implemented to facilitate data retrieval, organization, and archiving. File naming conventions and directory structures will be established to ensure ease of access and searchability.

2.4.4. Access Controls and Encryption

Access to the dataset will be tightly controlled and restricted to authorized personnel only. Role-based access control mechanisms will be employed to manage who can access and modify the data. Data will be encrypted both at rest and during transit to protect against unauthorized access or data breaches.

2.4.5. Data Retention Policy

A data retention policy will be established to define how long the data will be stored. It will comply with relevant data protection regulations. Data that is no longer needed for the project's objectives will be securely deleted or archived according to the retention policy.

2.5. Data Security and Privacy (specific)

This section discusses the steps undertaken to ensure data security and privacy. Before uploading the dataset, we need to address any personally identifiable information (PII) concerns in the dataset and explain how these have been anonymized or pseudonymized in the data, if necessary. This information complements the access controls and encryption methods to safeguard the dataset.

To ensure data security and protection, the following measures need to be implemented:

2.5.1. Anonymization of Sensitive Data

Personally identifiable information (PII) (e.g., license plates or faces) will be anonymized to prevent the identification of individuals or vehicles. Anonymization techniques may include blurring or obfuscation of sensitive information while preserving the utility of the data for research purposes.

2.6. Data Access and Sharing (common)

This section describes how the dataset is shared with others, such as project researchers or the public. It describes the access policies and guidelines for sharing the data, including any necessary data usage agreements or licenses.

As access control measures are crucial to protect the confidentiality, integrity, and availability of the project's data, the following security measures will be implemented to

ensure appropriate access and protect sensitive data: Complete datasets will be available for the project participants, and data-subsets will be made publicly available, and different data-access measures will be implemented for both groups.

2.6.1. Access measures for project participants

VPN Access:
Access to the data-storage system and its resources is restricted to authorized individuals only. A Virtual Private Network (VPN), managed by project members, is used to establish secure connections between remote users and the system. VPN access will require authentication using appropriate credentials and encryption protocols. Access logs will be maintained to monitor VPN connections and identify any suspicious activity.

Role-based Access Control (RBAC):
Access to different levels of data and system resources will be granted based on roles and responsibilities within the project. RBAC will be regularly reviewed and updated to ensure access privileges are current and appropriate.

User Authentication:
The system employs user and password authentication mechanisms to control access to the data. Each user will have a unique username and a strong, securely stored password using the SHA-2 encryption protocol.

2.6.2. Access measures for publicly available data

Publicly accessible data will be made available through a dedicated website. The following measures will be implemented to manage access to the public data,

User Registration:
Interested users will be required to register on the website to access the public data. Registration will involve providing necessary institution and contact information and agreeing to the terms and conditions of data usage. Users will be required to provide valid contact information to receive their login credentials.

User Agreement:
Prior to receiving access credentials, users will be required to sign an agreement outlining the terms of data usage. The agreement will specify the permitted uses of the data, any restrictions, and the user's responsibilities regarding data handling and confidentiality. The agreement will emphasize the need to comply with relevant laws, regulations, and ethical guidelines.

User Authentication:
Upon successful registration and agreement signing, users will receive login credentials (username and password) via email. Users will be required to authenticate themselves using their login credentials to access the public data.

Access Monitoring and auditing:
User accounts and activity will be monitored to detect and prevent unauthorized access or suspicious behaviour. System administrators will regularly review access logs to identify any potential security incidents. User activities within the website, including data downloads and queries, will be logged to enable auditing and monitoring. Auditing will help ensure compliance with the terms of data usage and investigate any misuse or unauthorized access.

Support and Assistance:
User support will be provided to address inquiries, troubleshoot technical issues, and assist with access-related matters through a dedicated email address.

2.7. Data Preservation (common)

This section outlines the plans for long-term data preservation. It explains how the dataset's usability and integrity is to be maintained over time, including any version control strategies or data curation practices. Along the project, we consider sharing the dataset with appropriate repositories or archival platforms to ensure its long-term availability.

Collected data will be preserved beyond the duration of the project for long-term accessibility to enable future research. This section outlines the strategies and measures that will be employed to preserve the data after the project concludes.

2.7.1. Archival Platform

Trusted data repositories or archival platforms that align with recognized preservation standards and practices will be identified and selected based on their long-term commitment to data preservation, sustainability, and accessibility, and that ensures redundancy and protection against data loss. Periodically, the selected repository or archival platform will be reassessed for ongoing suitability and migration to alternative platforms will be considered if necessary.

2.7.2. Version Control and Data Curation

Version control practices will be implemented to manage changes and updates to the data over time, including modifications to data, metadata, and data format conversions. A record of previous dataset versions will be maintained, ensuring that older versions are accessible for reproducibility purposes.

2.8. Ethical Considerations (specific)

This section needs to discuss any ethical considerations related to the collection and use of the dataset. Address issues such as bias, privacy, and potential social implications of the data. Consider obtaining necessary permissions or consents, especially when capturing data in public spaces.

3. Data Management: Dataset … (tbc)

References

[1] Authors, Title, Journal, Vol(Num):pp-pp, year. (DOI url)
[2] Authors, Title, in Proc of Conference, year. (DOI url)