**Names**: Oren Raz, Lam Kai Yi Karen

**Assignment**: REI603M AI Lifecycle – Assignment 2: Dataset Exploration and Analysis

**Brief Introduction to Project – Sign Language Speech to Text**

In this project we train a utilize computer vision algorithms and train a model to convert signed speech to text, it could be a prototype for a model that can bridge the gap in communication between a deaf and a hearing person (with no knowledge of sign language). Currently, there have been models capable of [classifying images of signed alphabets to its corresponding letter](https://huggingface.co/RavenOnur/Sign-Language). There have been [attempts to increase the accuracy of translation of signed speech](https://www.sciencedirect.com/science/article/pii/S2667305323001096) with the use of various image and video processing techniques.

The translation of sign language is a challenging task due to the vast amount of vocabulary used in daily conversations. The meanings of signs are also dependent on a combination of body motions, hands movements and head poses, sometimes even face expressions, which adds to the complexity of the project. Within the time frame of this module, we aim to develop a model to translate simple signed words or phrases into English phrases with greater accuracy.

**The Dataset – Word-Level American Sign Language (WLASL) Dataset**

**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 used in this project is “[A large-scale dataset for Word-Level American Sign Language (WLASL)](https://github.com/dxli94/WLASL)” by Dongxu Li, Cristian Rodriguez-Opazo, Xin Yu and Hongdong Li. This dataset was collected from many small datasets, to facilitate the development and evaluation of deep learning-based methods for word-level ASL recognition.

**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 research paper, which created this dataset, is supported by Australia Re- search Council ARC Centre of Excellence for Robotics Vision (CE140100016), ARC-Discovery (DP 190102261) and ARC-LIEF (190100080).

**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 is a short video, about 2 seconds each, few are over 8 seconds. Each instance is also labelled with an associated word, frame per second, bounding box, video id, source, url, and signer id.

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

There are 11980 videos for 2000 words. Ranging from 2 to 16 videos per word.

**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)?**

The dataset contains a small subset of the English vocabulary, therefore not practical for a real-life application. The original dataset is a combination of multiple datasets of signed words. However, the dataset we would be working on excludes 9103 missing videos, as URL for some videos became invalid overtime. This has not resulted in the loss of any words.

**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 instance is a video of a person gesturing an English word. Accompanied with information about the video, such as the word, the signer, the focus area for the video (bbox), the source for refences and some data about the video format itself.

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

Yes, a word. Perhaps, if time allows, we can record videos of ourselves gesturing whole sentences using words from the train set, for a more sophisticated testing.

**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.**

As mentioned above, no words are missing. However, certain videos cannot be downloaded due to invalid URL links. There is still a significant number of videos to analyse, excluding those that are missing. Therefore, the model can be trained on the available videos.

**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.**

The table with information about the videos indeed has a column about the split, however, during our data exploration process we found out that this split is not relevant for us, possibly due to the missing videos. We will split it better to ensued isolating some signers from the train set, while keeping in mind fair word distribution: no new words in test and validation, no wasted words in train without validating or testing them.

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

Not as we know according to the data collection paper. But since there are a few exceptionally long videos (over 8 seconds) we suspect there might be errors, we will check whether they contain more than one word.

**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 dataset consumer? 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 relies on videos from multiple websites that are downloaded by a provided script. The script logs each id of failure and writes to a file. The file containing the missing data ids can be uploaded to an online form to request the missing data from the author. So far, we have not got a response from them and will work with what we have.

**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’ nonpublic communications)? If so, please provide a description.**

No, this data is public and available for research and education, non commercial.

**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.**

The faces of the signers are shown, and each is given a unique signer id only for this dataset (not SSN or something like that). Image recognition models might identify the people, but they recorded themselves while aware of the exposure.

**Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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.**

Apart for the signers faces, no info about them is shared in this dataset.

**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.**

The resolution of all original video frames was resized such that the diagonal size of the person bounding box is 256 pixels.

**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.**

Not mentioned, but we assume that the authors kept it, however we don’t have it.

**Uses**

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

Yes, the authors collected it specifically to train a model, then other machine learning enthusiast tried to train their own models. [Here](https://github.com/alanjeremiah/WLASL-Recognition-and-Translation) is one example of such attempt.

**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.**

Can be found [here](https://github.com/dxli94/WLASL/blob/master/README.md).

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

Face expression sentiment analysis, some signs combine hands movement and face expression.

**Distribution** (Not relevant)

**Maintenance**

**Who will be supporting/hosting/maintaining the dataset?**

No one, however the author made it possible to apply for access to missing data, deadline for this service is not mentioned

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

Their emails appear on the paper, however these are university address and might be unused, a link to an access request form to the full dataset is available in the repo.

**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 dataset consumers (e.g., mailing list, GitHub)?**

Not as we know, but if we find an error we update it for ourselves.

**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 dataset consumers? If so, please provide a description.**

Contribution mechanism is not mentioned, however one can try to open a pull request and hope for it to be approved.