Student: Elshan Naghizade

Project: Python Library To Transform Image Datasets in Query-able SQL Tables

Date: 25 June, 2023

RESEARCH STRATEGY

DEVELOPMENT STRATEGY	QUANTITATIVE ANALYSIS STRATEGY
Data Extraction: I need to construct functions capable of reading image datasets from a myriad of formats such as JPEG, PNG, TIFF, BMP, etc. Utilizing libraries like PIL (Pillow) or OpenCV would facilitate this task. The functions ought to handle both singular images and directories replete with subdirectories of images. **Feature Extraction:** Upon successful loading of the images, the next phase involves extracting meaningful features. This could range from basic features like color histograms, texture, shape to more complex ones like SIFT, SURF, or deep learning features derived from pre-trained models. OpenCV is the go-to for simple image features, whereas scikit-image caters to advanced features. **Data Transformation:** Once features are extracted, they need to be converted into a query-able format such as SQL. I choosing leveraging libraries like SQLAlchemy or sqlite3 to spawn SQL tables and insert data into them.	**Memory Usage:** The goal here is to measure how much memory the library uses since it mostly operates in RAM. Python's 'memory-profiles' library is to be used as the main benchmark. **Scalability:** This involves analyzing how the library's performance scales with increasing dataset size. It's critical to understand how the tool performs when processing small, medium, and larger datasets, and whether it maintains its speed proportionally as the data volume grows. Python's 'time' library will be used to log the corresponding timestamps. **Flexibility:** This metric could involve measuring the variety of image formats and feature types the library can handle compared to manual methods.

DATA ACQUISITION STRATEGY

I will need any 3 image datasets of different sizes to benchmark its scalability. So simply sorting by dataset size on Kaggle does the job.

1) Small (9 Mb) - CAPTCHA Images: https://www.kaggle.com/datasets/fournierp/captcha-version-2-images
2) Medium (55 Mb) - Celebrity Face Image Dataset: https://www.kaggle.com/datasets/vishesh1412/celebrity-face-image-dataset
3) Large (264 Mb) - WeedCrop Image Dataset: https://www.kaggle.com/datasets/vinayakshanawad/weedcrop-image-dataset

DATA CLEANSING

The library is to be used in lieu of data loaders so the actual cleansing of datasets is out of the scope of my package.