ALaBDac - Automated Lab Book Data Collection

Problem Overview

Researchers often take meticulous notes of the work they do in the lab, which is often written into lab books. The downside is, of course, that lab books are hard to share and search at a later date. This project aims to develop an automated way to collect lab book information and automatically record it to a central data store as its written.

Objectives:

- Image Capture - take a photo of an unobstructed lab book
- Keyword extraction - Extract keywords from image
- Storing the data - so it can be accessed

Key Hardware:

- Jetson Nano
- Raspberry Pi High Quality Camera
- Camera Lens 6mm Wide Angle
- RFID Reader

Key Packages

This system is written in Python with a few major packages used:

- Image manipulation - OpenCV, Numpy, Pillow
- Machine Learning - Pytorch
- Data storage - Boto3, LMBD

Unsupervised Image Capture

An unsupervised method was used, as a dataset of hands wearing scientific clothing with a lab book background could not be sourced. Thus a coloured border was put around the lab book, and the amount of the border in the frame is calculated. The difference between frames is used to check if there is activity happening in the frame. Frame stacking is applied to improve the signal to noise ratio and smooth defects in the images.

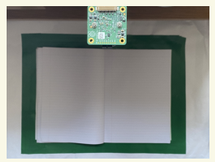

Figure: Example border around lab book

Keyword Extraction Machine Learning Pipeline

As there is not a pre-existing handwritten dataset for all words in the English language, the human handwriting synthesizer ScrabbleGAN was used. ScrabbleGAN is used to generate the keywords to be detected on a lab book page; some example words generated using ScrabbleGan are shown in the figure below:

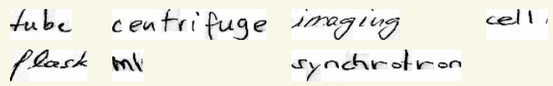

Figure: ScrabbleGAN generated words (tube, centrifuge, imaging, cell, flask, ml and synchrotron)

These words are manipulated (resized, background removed, greyed) and then superimposed on lab book images to create the object detection algorithm's training and testing dataset. This pipeline allows for an object detection model to be trained to identify a set of keywords on a lab book background.

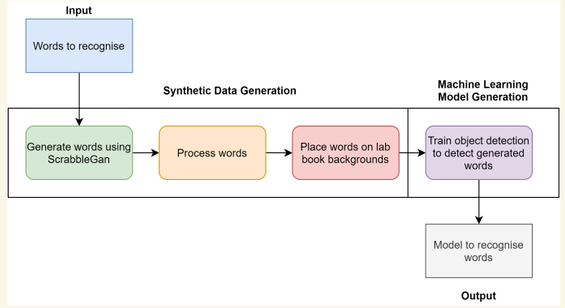

Figure: Keyword extraction machine learning pipeline

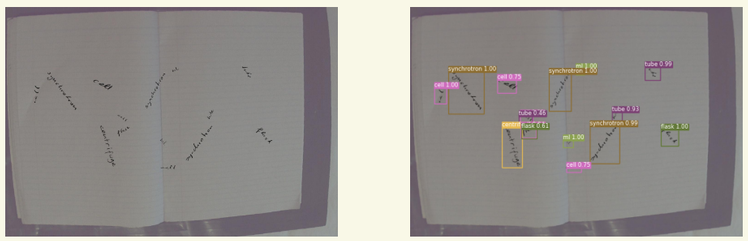

Figure: Example detection on synthetic test data (input, output)

Storing Data

The images will be stored in an S3 bucket, as objects with the keywords being placed in the metadata of the image, as demonstrated in figure below.

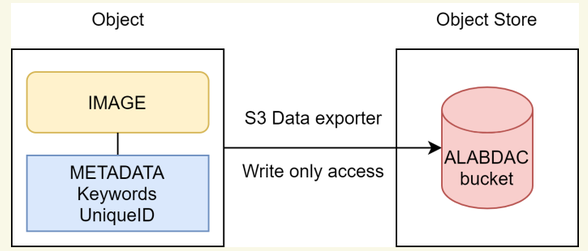

Figure: S3 Data exporter

The System

The system will be started and stopped with an RFID card being read. Once started, it will initialise, and it will start running; all processing will be run on the Jetson Nano, with the data stored in a bucket.

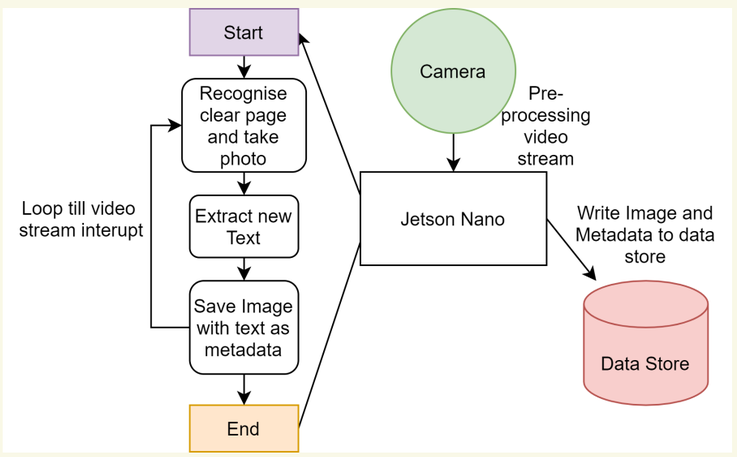

Figure: System diagram

Acknowledgments

I would like to thank my supervisors Dr David Walker (david.walker@plymouth.ac.uk), Dr Mark Basham (mark.basham@rfi.ac.uk) and Dr Laura Shemilt (laura.shemilt@rfi.ac.uk) for help and useful guidance on this project.

Gabryel Mason-Williams
gabryel.mason-williams@students.plymouth.ac.uk
University of Plymouth