Pawsey Supercomputing Centre
Research Internships
2017-2018

Intern Project Proposal

PROJECT DETAILS

Project Title	Historical Image Classification System

Project Background:
Historical image collections such as that held by the Battye Library of West Australian History at the State Library of Western Australia (SLWA) hold enormous value for their documentation of cultural and built development. Large image collections can be hard to interrogate because searching is usually reliant on text-based search of manually entered meta-data, and the quality of the results is reliant on the search terms used and the nature of the meta-data that has been entered. The use of image processing, machine learning and computer based methods to semi-automatically categorise and classify images offers new opportunities to interpret large image collections and gain added value out of such collections.

The SLWA has a current catalogue holding of approximately 700,000 images and receives donations of many thousands of images each year. Many of these images are historical photographs and as such have none of the modern metadata (often stored by the camera) such as date taken, camera details, exact location, etc. The sheer size of this collection makes the process of entering and creating such metadata a huge job.

A previous Pawsey project used a technique called feature matching (using SIFT) to perform a visual match on a collection of ten thousand images on Magnus. From this data a basic clustering was performed based on the number of matching points between images to help identify matching images, potential panoramas and multiple photographs of the same object – which in turn could be used for photogrammetric 3D reconstruction purposes. This project featured in a story on Channel 7's Today Tonight in March 2017. While it is possible to perform this technique on a sizable collection of images it is impractical for a very larger collection due to the amount of compute required. To reduce the amount of compute it would be necessary to use metadata and other classification information to make the processing more manageable.

Project Proposal:
The aim of this project is to use machine learning and image processing to assist with classifying images and automating the metadata generation and image clustering processes. This would have several benefits ranging from assisting with supplementing existing metadata, generating new metadata for existing images, as well as potentially assisting with creating the initial metadata for new images to be included in the catalogue.

Some image processing steps that could be performed include:

- Face detection and possibly recognition/matching (possibly using OpenCV)

- Detection and optical character recognition (OCR) of text

- Detection of human recognisable features such as buildings, streets, signs, etc.

- feature matching to only use larger scale features

- Image clustering based on the features and processing described above

This project would interest a student with a background in programming and interests in image processing, history and supercomputing.

References:

- OpenCV: http://opencv.org/

- SIFT: https://en.wikipedia.org/wiki/Scale-invariant_feature_transform

- Face detection with OpenCV:
http://docs.opencv.org/3.1.0/d7/d8b/tutorial_py_face_detection.html

- Chris Norman "Feature Matching the State Library of Western Australia's Photographic Archive" Pawsey student project report, March 2017.

Primary Supervisor/Institution	Joshua Hollick / Curtin University
Contact Details	Joshua.Hollick@curtin.edu.au

What makes this a Supercomputing Project?
Training of image classifiers, neural networks and scale of compute for large datasets

What is the application/code?	We are looking at using some of our custom code developed for the Sydney/Kormoran project as well as OpenCV and various other image processing applications.
Pawsey Resources to be Used	Supercomputing and Nimbus Research Cloud
How many core-hours are required?	100000
STUDENT ATTRIBUTES	
Academic Background	Computer Science, Image Processing or Software Engineering Ideally with programming skills
Expected Computing Skills	C++, Python or other suitable language for implementing image classification. Git version control. Ideally familiar with OpenCV or similar.
Additional Pawsey Training Requirement	OpenMP

PROJECT TIMELINE (DRAFT)	
Week 1	*Pawsey Induction – 27 Nov - 1 Dec 2017*
Week 2	HIVE Induction, Familiarise with SLWA Catalogue, Project planning
Week 3	Familiarise with photogrammetry, OpenCV, image classification, machine learning
Week 4	Code Development and small testing
Week 5	Code Development and small testing

Week 6	Run classification on subset of data and test results
Week 7	Review results to date and where further time should be spent
Week 8	Larger testing
Week 9	Buffer
Week 10	Prepare report, presentations and other outputs
Final Presentation	*Pawsey Internship showcase - TBD February 2018*