Training Report – ASTRI Summer Internship Programme 2021

PARTICULARS OF SUMMER INTERN

Office/TD/CCG CTO/CCT/MSA

Name of Summer Intern (in English) Daniel Shiu-hong Kao Graduation Year 2023

Summer Internship Period From Jun. 28, 2021 To Aug 20, 2021

Name of Mentor Bob Koon

Purpose of Training Report

This report is part of the summer internship programme to record the work completed done and learning experience of the summer intern which enable us to make the assessment process more manageable for Summer Intern and Mentor.

Guidelines

The Summer Intern should prepare training report including but not limited to the following components:

(a) Work Completed
(b) Technical knowledge / project involvement
(c) Problems encountered
(d) Contributions / Learnings

Submission timeline

Summer Intern:

(a) Receive your training report around midway of your internship from Talent Acquisition team via email.
(b) Submit your training report to your mentor by **the second last week of your internship**.

Mentor:

(a) Review and complete the training report by **the last week of the internship**.
(b) Send the completed and signed training report to executive assistant. If your team does not have an executive assistant, please send it to Talent Acquisition Team directly.
(c) Executive assistant will send it back to Talent Acquisition Team.

Part 1: TRAINING REPORT

Summary of work undertaken during the internship, highlighting what you have **observed and learned.**

I’ve been assisting my mentor as well as his teammates to do some projects. Projects included:

(1) **C++ OCR development using Tesseract:**

Time period: 1st – mid-2nd week of internship

Task assigned by: Bob

Content: Improve the Tesseract performance of date recognition in HKID card, including date-of-birth and issue date.

Result: In the testing data set, the accuracy of recognition of issue date was raised from 85.19% to 100.00%, while of date-of-birth was improved from 90.07% to 98.07%.

Resource: Tesseract - https://tesseract-ocr.github.io/

(2) **C/C++ Libzip development for zipped files containing text files and repositories:**

Time period: mid-2nd – mid-3rd week of internship

Task assigned by: Bob

Content: Using Libzip in C/C++ to develop methods for zipped files compression and extraction, as well as reading the content of each text file inside.

Result: Methods for compression and extraction are done. At the same time, two other methods for building zipped file based on a series of input string and reading text file inside a zipped file are both achieved.

Resource: Zlib - https://github.com/kiyolee/zlib-win-build,
Libzip - https://github.com/kiyolee/libzip-win-build

(3) **SVG development and Makemeahanzi testing for Chinese character stroke animation:**

Time period: mid-3rd – mid-4th week of internship

Task assigned by: Bob

Content: Navigating the SVG data set in Makemeahanzi, and try to form a Chinese character with strokes, which is an array of points.

Result: Lack of common point in the intersection among different stokes leads to an uncertainty of the goal availability. Task postponed in the end.

Resource: Makemeahanzi - https://github.com/skishore/makemeahanzi

(4) **Dataset building and labeling:**

Time period: mid-4th – mid-5th week of internship

Task assigned by: Adrian

Content: Helping build and label a dataset for typed and hand-written English characters.

Result: Task done.

Resource: None

(5) **KSAI - Simplified Chinese OCR library testing:**

Time period: 5th week of internship

Task assigned by: Bob, Vincent

Content: Build the environment for KSAI and test the output in different kinds of situations. Also, interpret the architecture of each model in use and understand the output meaning for each model.

Result: KSAI supports simplified Chinese and English characters OCR. However, it has good performance only for typed characters in horizontal direction. Besides, CNN and RNN are used in the model. There are three models in total. One is for box detection; another is for word direction; the others work as single line characters recognition.

Resource: https://github.com/kingsoft-wps/KSAI-Toolkits

(6) **Microsoft Azure Form Recognizer testing:**

Time period: 5th week of internship

Task assigned by: Vincent

Content: Test the performance of Azure Form Recognizer on an application form from ASTRI.

Result: I could use the client model provided by Microsoft successfully. The performance is acceptable but not perfect. Maybe one can try to use a custom model in the future to get a better accuracy.

Resource: https://azure.microsoft.com/en-us/services/form-recognizer/

(7) **Parsing of custom input of Hong Kong residential address:**

Time period: early-6th week of internship

Task assigned by: Bob

Content: Parsing the room/flat, floor, block information in the provided dataset of Hong Kong Address, where the input may contain some noises.

Result: I used C++ to parse the input, and the accuracy was about 85%.

(8) **Improvement for Hong Kong address parsed from the Internet:**

Time period: mid-6th week of internship

Task assigned by: Bob

Content: making a web scraping code to parse the Hong Kong residential address from Midland. Afterwards, generalizing all the possible patterns for room and floor information.

Result: I wrote a program in python, using Beautifulsoup, to grab Hong Kong address from Midland. The address should include region, district, street, street number, estate, block, floor, and unit. Then, I generalize the possible titles for floor and unit.

(9) **Image blurring experiment with content maintenance:**

Time period: 7th week of internship

Task assigned by: Bob

Content: Giving two images with the same size. One is the original photo, and the other shows the part of character inside. Try to produce a image with those characters in the same color, and blur the background.

Result: Task done, and also we have another method to test the blurriness of a specific image.

Daniel Kao

Aug 5, 2021

Signature: Date:

Name of Summer Intern: Daniel Shiu-hong Kao

Part 2: MENTOR'S COMMENTS:

Daniel shows his talent and professional during his internship period. He has designed and developed algorithm for the projects which solve the problem from the mentor. He also helps on studying latest technology and make the report on what he found.

5-8-2021

Signature: Date:

Name of Mentor: