MSc Individual Project Progress Report

Michael Song (02405765) on 28 April

Task Description

The project *Natural Language Processing: Applied to Modern Literature* include two main tasks:

1. Optical Character Recognition (OCR) on scanned Chinese journal, which consists of over 100 articles. Transfer the image of these articles to text data and construct a digital database.
2. Apply Natural Language Processing (NLP) techniques to analyze the text data, extract information on gender discourse, i.e expectations of women, as well as attitudes towards gender relations.

Task 1 is the main goal of my research. The challenges include:

- Most of the Chinese characters on the image are low in resolution (under 20*20 pixels), making them hard to distinguish.
- The quality of the image varies (noise, blur, lighting, etc.).
- The images have to be downloaded one at a time. Additional script might help with batch downloading.
- Some ready-to-use OCR tools have a poor performance on the article images.

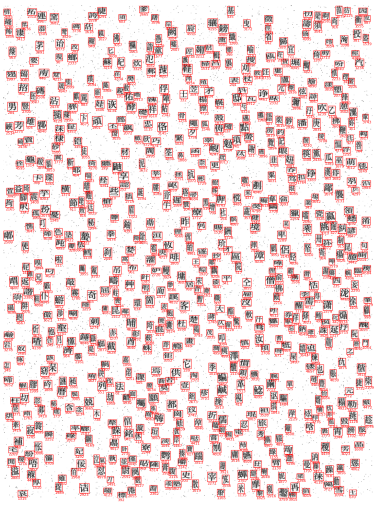
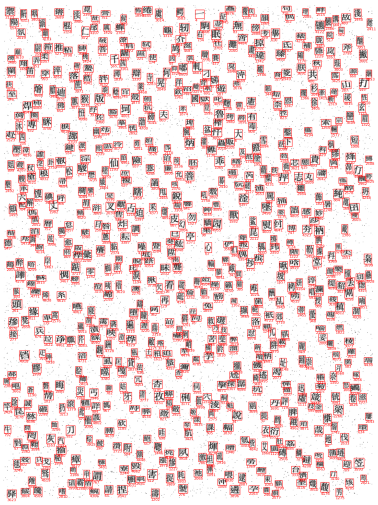
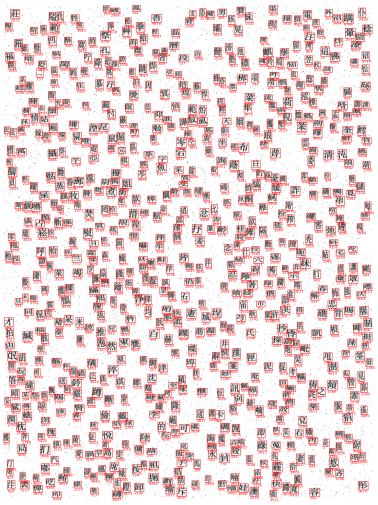
Progress

The current achievements include:

- Literature review of OCR techniques, with a focus on Chinese characters (20+ publications).
- Experiment with some popular model architectures like *ResNet* and *Faster-RCNN* for character recognition/classification.

The methods to collect training data are:

1. Use 5000 common Chinese characters to create an image dataset D1 (one character per image).
2. Use dataset D1 to create synthetic image dataset D2, with multiple characters, their labels and bounding boxes in each image (as shown below, red bounding boxes and labels are for visualization only).

The methods to train the models are:

1. Train a character **classifier** (*ResNet*) using dataset D1.
2. Train a character **detector** (*Faster-RCNN*) using dataset D2 (only predict position).
3. Train a character **recognizer** (*Faster-RCNN*) using dataset D2 (predict position and label).

The experiment conducted is to compare the performance of the following two methods:

1. Use the **detector** to extract image patches which contains a single Chinese character, and use the **classifier** to predict its label (need to train two models separately).
2. Use the **recognizer** to predict character position and label at the same time (end-to-end training).

The experiment is not finished now.

Future Plans

The timeline of the future plan is flexible, which includes:

1. Finish the current experiment.
2. For the better approach in the experiment, improve the performance by refining model structure, training pipeline etc.
3. Write programs to auto-download the target images in the website.
4. Transfer the image of these articles to text data using the proposed method.
5. Compare the performance of the proposed method with SOTA methods on the target images.
6. Finish Background and Progress Report (04 Jun), Final Report (09 Sep), and Presentation Slides (13 Sep).

Questions

1. Regarding the whole project, since Task 2 (NLP) depends on the OCR result of Task 1, what is the expected deadline for me to finish Task 1?
2. What **kind** of article images in the website do I need to process? (categories include advertisement, drawings, discussions, learning, statement, housekeeping, masterpiece, novels etc.)
3. What **period** of article images in the website do I need to process? (available period is from 01/1915 to 12/1931)
4. What is the expected OCR accuracy on article images?
5. What is your preferred way of discussing the project? This is not limited to:
   - Regular progress report via email (like this)
   - Online or in-person meetings
   - WeChat group/individual messages