MSc Individual Project Progress Report

Michael Song (02405765) on 27 July

Progress before last meeting

- Background and progress report
- Data collection (journal images with labels)
- Synthetic datasets for training
- Trained following models:
  - Text Line Detector (LD), based on Faster-RCNN
  - Text Line Recognizer (LR), based on CRNN
  - Character Detector (CD), based on Faster-RCNN
  - Character Recognizer (CR), based on Faster-RCNN
  - Character Classifier (CC), based on ResNet
- Compared the performance with following methods:
  - CR
  - CD + CC
  - LD + CR
  - LD + CD + CC
  - LD + LR
- **LD + LR** is better and chosen for further improvement
  - LD performs well on the most common layout.
  - LR only performs well on train/validation set, **but not** on target images.
    - The synthetic training set may not fully matches the target data.
  - Both LD and LR are only trained on vertical text images.
- Manual text extraction performed on 20 images as **testing set**.
- **LD** Model Training:
  - Synthetic dataset with random layouts and icon integration.
  - Extensive data augmentation techniques used.
- **LR** Model Training:
  - Combined vertical and horizontal text, frequency-based character selection.
  - Typeface variety increased from 1 to 4.
  - Extensive data augmentation applied.
- Results:
  - LD and LR can handle diverse text layouts.
  - LD is better at ignoring non-text areas.
  - LR shows reduced over-fitting.
- Metrics:
  - Inference speed: **0.8** sec/image.
  - CER: **0.44**
  - BLEU-4: **0.41**
- Comparison with Baseline:
  - Outperforms Apple's OCR engine in CER and BLEU-4.
- Processed textual data from 35,851 journal images.

Progress since last meeting

Training

- Typeface variety increased from 4 to 12
- Wider range of data augmentation
- LR model architecture adjustment
- Introduce **word-based** character selection in training data:
  - Insert common Chinese words from pretrained word vectors (data source) into the training set.
  - This makes the model "memorize" meaningful words.
  - This can be used together with **frequency-based** character selection.

Result

- Average CER improved to **0.33** for 20 images
  - **0.28** for 17 common layout images
  - Less than **0.2** on half of the images
- Average BLEU-4 improved to **0.55** for 20 images
  - **0.62** for 17 common layout images

User Interface

- Load and display image locally
- Zoom and shift the image
- Perform OCR on the image
- Show the results of both detection and recognition
- Switch between traditional and simplified Chinese

Future work

- Further model improvement
- Additional baseline model comparison
- Final report

Questions

- Change report title to "Optical Character Recognition: Applied to Modern Literature"?