UNIVERSITY OF TECHNOLOGY SYDNEY

Deep Learning-Based Text Detection and Recognition

by Qingqing Wang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Xiangjian He & Michael Blumenstein & Wenjing Jia

University of Technology Sydney
Faculty of Engineering and Information Technology

April, 2020

Certificate of Authorship/Originality

I, Qingqing Wang, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with East China Normal University.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 7/4/2020

Acknowledgements

I would like to acknowledge all people and institutions that have provided support during my PhD candidature.

First of all, I thank my supervisors Prof. Yue Lu, Prof. Xiangjian He and Prof. Michael Blumenstein for their patient advice and guidance in my academic career. They are all knowledgeable, enthusiastic, optimistic, wise and helpful people. I have learned so much from them, not only how to be a successful scholar but also how to be a kind people in life. My supervisors have provided me such a precious opportunity to study overseas at the University of Technology Sydney (UTS). I never thought that such a cool thing would happen to me. Prof. Prof. Yue Lu is one of the most important people in my life. I jointed his research team in 2013 as a newbie in research. Since then, Prof. Yue Lu has taught me how to do research from scratch and led me to the right academic way. I know Prof. Yue Lu has always been doing his best for me and believe his knowledge, experience and vision enable him to help me do the best choice in every crossroad of my life. I am honoured to be supervised by Prof. Xiangjian He and Prof. Michael Blumenstein in UTS. Both of them are nice and humble in life and with very good reputation in our research community. Studying abroad alone is not easy, but with their help and support, I have enjoyed a wonderful time in Sydney.

I would also like to express gratitude to Dr. Wenjing Jia for her valuable comments on my writing skill. I believe what I have learned from her will benefit me for life. Additional thanks to students in my LPAIS lab and CVPR lab, including Wenqiao Sun, Hongjian Zhan, Xuecheng He, Yuhuang Xiu, Xiaohua Wei, Ye Huang, Xiaochen Fan, Yue Xi, Lei Liu, Saeed Amirgholipour Kasmani, Hesam Hesamian, etc. They have altogether created a wonderful research environment and made me enjoy my research in East China Normal University (ECNU) and UTS. The days of pursuing a dual PhD degree are sometimes full of darkness, loneliness and depression. Thank my friends Qin Chen, Ye Huang and others for accompanying and supporting me during these tough times.

Next, I am grateful for the financial supports from ECNU, UTS, Chinese Scholarship Council (CSC) and BYKER DIGITAL BIOTECHNOLOGY CO., LTD. They make me relaxed on my way to dream.

Finally, I sincerely thank my parents and sisters for their love, understanding and encouragement in this long journey. Their support is always my strength and lets me feel free to pursuit my dream. Being a member of such a wonderful family is the most beautiful thing that happened to me. I am so proud of you all and thank you very much for everything you have done for me.

Qingqing Wang

Sydney, Australia, 2020.

List of Publications

Published Papers

1. **Qingqing Wang**, Ye Huang, Wenjing Jia, Xiangjian He, Michael Blumenstein, Shujing Lyu and Yue Lu. FACLSTM: ConvLSTM with focused attention for scene text recognition. SCIENCE CHINA Information Science, 2020, 63(2): 120103. (JCR: Q1)

2. **Qingqing Wang**, Wenjing Jia, Xiangjian He, Yue Lu, Michael Blumenstein, Ye Huang and Shujing Lyu. DeepText: detecting text from the wild with multi-ASPP-assembled DeepLab. *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 208-213, 2019. (Top: A)

3. **Qingqing Wang**, Wenjing Jia, Xiangjian He, Yue Lu, Michael Blumenstein, Ye Huang and Shujing Lyu. ReELFA: a scene text recognizer with encoded location and focused attention. *International Conference on Document Analysis and Recognition Workshops(ICDARW)*, pp. 71-76, 2019. (Top: A)

4. Yuhuan Xiu, **Qingqing Wang**, Hongjian Zhan, Yue Lu and Man Lan. A Handwritten Chinese Text Recognizer Applying Multi-level Multimodal Fusion Network. *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1464-1469, 2019. (Top: A)

5. **Qingqing Wang** and Yue Lu. A sequence labeling convolutional network and its application on handwritten string recognition. *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2950-2956, 2017. (Top: A*)

6. **Qingqing Wang** and Yue Lu. Similar handwritten Chinese character recognition using hierarchical CNN model. *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 603-608, 2017. (Top A)

7. **Qingqing Wang**, Xiao Tu, Shujing Lu and Yue Lu. Text extraction from mail images with complex background. *International Forum of Digital TV and Wireless Multimedia Communication (IFTC)*, pp. 3-11, 2017.

8. Hongjian Zhan, **Qingqing Wang** and Yue Lu. Handwritten digit string recognition by combination of ResNet and RNN-CTC. *International Conference on Neural Information Processing (ICONIP)*, pp. 583-591, 2017. (Top A)

9. Ruyu Zhang, **Qingqing Wang** and Yue Lu. Combination of ResNet and center loss based metric learning for handwritten Chinese character recognition. *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 25-29, 2017. (Top A)

10. Shangxuan Tian, Ujjwal Bhattacharya, Shijian Lu, Bolan Su, **Qingqing Wang**, Xiaohua Wei, Yue Lu and Chew Lim Tan. Multilingual scene character recognition with co-occurrence of histogram of oriented gradients. Pattern Recognition, 51(2016), pp. 125-134. (JCR: Q1)

11. **Qingqing Wang**, Yue Lu and Ying Wen. Scene text detection using sequential nontext filtering. *International Conference on Image Processing (ICIP)*, pp. 1742-1746, 2015. (Top A)

12. **Qingqing Wang**, Yue Lu and Shiliang Sun. Text detection in nature scene images using two-stage nontext filtering. *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 106-110, 2015. (Top A)

Contents

List of Figures

List of Tables

Abbreviation

CC - Connected Component

MSER - Maximal Stable Extremal Region

LSTM - Long Short-Term Memory

CNN - Convolutional Neural Network

FC-LSTM - Fully-connected-LSTM

SPP - Spatial Pyramid Pooling

ASPP - Atrous Spatial Pyramid Pooling

IoU - Intersection-over-Union

BN - Batch Normalization

RNN - Recurrent Neural Network

ReELFA - Recognizer with Encoded Location and Focused Attention

ConvLSTM - Convolution LSTM

FACLSTM - ConvLSTM with Focused Attention

HMM - Hidden Markov Model

ICDAR - International Conference on Document Analysis and Recognition

HOG - Histogram of Oriented Gradients

LBP - Local Binary Pattern

SWT - Stroke Width Transform

ERs - Extremal Regions

SVM - Support Vector Machine

RF - Random Forests

FCN - Fully Convolution Network

SSD - Single Shot Multibox Detector

RPN - Region Proposal Network

NMS - Non-maximum Suppression

RoI - Region of Interest

FPN - Feature Pyramid Networks

TCM - Text Context Module

RRPN - Rotated Region Proposal Network

CTPN - Connectionist Text Proposal Network

ITN - Instance Transformation Network

CTC - Connectionist Temporal Classification

STN - Spatial Transformer Network

EP - Edit Probability

HAM - Hierarchical Attention Mechanism

TPS - Thin-Plate-Spline

HTR - Handwritten Text Recognition

Semi-CRFs - Semi-Markov Conditional Random Fields

FNNLM - Feedforward Neural Network Language Model

RNNLM - Recurrent Neural Network Language Model

MDLSTM - Multi-Dimensional Long-Short Term Memory

PCA - Principal Components Analysis

WFST - Weighted Finite-State Transducer

HCR - Handwritten Character Recognition

MQDF - Modified Quadratic Discriminant Function

ATRCNN - Alternately Trained Relaxation Convolutional Neural Network

DirectMap - Direction-decomposed Feature Map

AP - Average Precision

MST - Minimum Spanning Trees

CRF - Conditional Random Field

SVHN - Street View House Number

NAS - Neural Network Search

PAN - Pyramid Attention Network

PMTD - Pyramid Mask Text Detector

ABSTRACT

Deep Learning-Based Text Detection and Recognition

by

Qingqing Wang

Texts play a critical role in our daily life. They are everywhere such as slogans on posters, licence plates on cars, etc., to transmit information and knowledge. With the popularity of mobile devices with cameras, more and more texts are collected, transmitted and stored as text images. Automatically reading texts from images is of high application potentials. Therefore, related researches have been attracting considerable attentions from the computer vision community. Scene texts and handwritten texts are the two most difficult texts to be automatically read because of the challenges posed by the complexity of backgrounds, the uncertainty of capturing conditions, the diversity of text appearances, touching characters and the variety of handwriting styles.

Text detection, *i.e.,* localizing text areas from images, and text recognition, *i.e.,* transcribing located text areas into character sequences, are two key steps of robust text reading. In recent years, they have entered a deep learning era, where Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) play important roles. Here, we conduct researches on text detection and recognition based on CNN and LSTM, as presented below.

1. To improve the recall rate of small text areas in oriented text detection, we propose an Xception-based multi-ASPP-assembled scene text detector named DeepText. DeepText inserts multiple Atrous Spatial Pyramid Pooling (ASPP) modules into Xception after feature maps with different resolutions to retain richer information for small text areas, and introduces auxiliary connections and auxiliary losses to speed up convergence and boost the dis-

crimination ability of lower encoder layers.

2. To address the issue that Mask R-CNN cannot fully leverage global information when performing predictions, we propose a scene text detector named GMask R-CNN, where a global mask module is designed to perform semantic segmentation by considering global information.

3. To tackle the problem that LSTM neglects the valuable spatial and structural information of 2-D text images, we propose two scene text recognisers named FACLSTM, which exploits convolution LSTM to directly perform sequential transcription in 2-D space, and ReELFA, which utilizes one-hot encoded locations to enhance features with pixels' spatial information.

4. To solve the problem that CNNs with fully connected layers are not suitable for sequential prediction tasks due to their requirements of fixed-size inputs/outputs, we propose a CNN-based handwritten text recogniser CFSPP. CFSPP embeds a Spatial Pyramid Pooling-based intermediate layer between convolutional layers and fully connected layers to convert arbitrary-size feature maps into feature vectors with specific lengths.