

# **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

Certificate	ii
Acknowledgments	iii
List of Publications	v
List of Figures	xii
List of Tables	xvi
Abbreviation	xviii
Abstract	xx
<b>1 Introduction</b>	<b>1</b>
1.1 Background of Text Detection and Recognition . . . . .	1
1.2 Motivation . . . . .	9
1.3 Related Works . . . . .	11
1.3.1 Convolutional Neural Network and Long Short-term Memory for Text Detection and Recognition . . . . .	11
1.3.2 Text Detection . . . . .	15
1.3.2.1 Region-based and CC-based Detectors . . . . .	15
1.3.2.2 CNN-based Detectors . . . . .	18
1.3.3 Text Recognition . . . . .	24
1.3.3.1 Traditional Segmentation-based Recognisers . . . . .	25
1.3.3.2 Segmentation-free Recognisers Based on CNN and LSTM . . . . .	27

1.4	Evaluation Metrics . . . . .	32
1.4.1	Text Detection . . . . .	32
1.4.2	Text Recognition . . . . .	33
1.5	Contributions and Thesis Organisation . . . . .	34
<b>2</b>	<b>Multi-ASPP Assembled DeepLab for Multi-oriented Text Detection</b>	<b>37</b>
2.1	Introduction . . . . .	37
2.2	Proposed Method . . . . .	38
2.2.1	Backbone of Proposed DeepText . . . . .	40
2.2.2	Output Layer . . . . .	42
2.2.3	Multiple ASPP Layers . . . . .	42
2.2.4	Multiple Auxiliary Losses and Connections . . . . .	43
2.3	Experiments . . . . .	45
2.3.1	Datasets . . . . .	46
2.3.2	Implementation Details . . . . .	46
2.3.3	Evaluation of the Proposed Detector . . . . .	46
2.4	Conclusion . . . . .	48
<b>3</b>	<b>Mask R-CNN with Global Text Context for Multi- lingual Multi-oriented Text Detection</b>	<b>50</b>
3.1	Introduction . . . . .	50
3.2	Proposed Method . . . . .	52
3.2.1	Mask R-CNN . . . . .	53
3.2.2	Global Mask Module . . . . .	54
3.2.3	Loss Function of Proposed GMask R-CNN . . . . .	55



3.2.4	Data Augmentation and Configurations . . . . .	56
3.3	Experiments . . . . .	57
3.3.1	Datasets . . . . .	57
3.3.2	Implementation Details . . . . .	57
3.3.3	Comparison Results . . . . .	57
3.4	Conclusion . . . . .	60
<b>4</b>	<b>ConvLSTM-based Neural Network for Scene Text Recognition</b>	<b>61</b>
4.1	Introduction . . . . .	61
4.2	Proposed Method . . . . .	64
4.2.1	CNN-based Feature Extraction . . . . .	65
4.2.2	Sequential Transcription Module . . . . .	68
4.2.3	Training . . . . .	72
4.3	Experiments . . . . .	73
4.3.1	Datasets . . . . .	73
4.3.2	Implementation Details . . . . .	74
4.3.3	Experimental Results . . . . .	74
4.3.3.1	Comparison with Methods based on the Traditional FC-LSTM . . . . .	74
4.3.3.2	Comparison with Non-LSTM based Methods . . . .	76
4.3.3.3	Ablation Study . . . . .	77
4.4	Conclusion . . . . .	79
<b>5</b>	<b>FC-LSTM-based Neural Network for Scene Text Recognition</b>	<b>81</b>

5.1	Introduction . . . . .	81
5.2	Proposed Method . . . . .	82
5.2.1	One-hot Encoded Location . . . . .	83
5.2.2	Attention-LSTM-based Sequence Transcription . . . . .	84
5.3	Experiments . . . . .	85
5.3.1	Comparison with Other Scene Text Recognisers . . . . .	85
5.3.2	End-to-End Scene Text Reading System . . . . .	88
5.4	Conclusion . . . . .	89
<b>6</b>	<b>Flexible SPP for CNN-based LSTM-free Handwritten Text Recognition</b>	<b>90</b>
6.1	Introduction . . . . .	91
6.2	Proposed Method . . . . .	93
6.2.1	CNN-based Feature Extraction . . . . .	95
6.2.2	Reimplementation of SPP . . . . .	96
6.2.3	Flexible SPP Layer . . . . .	98
6.2.4	Model Prediction . . . . .	100
6.3	Experiments . . . . .	102
6.3.1	Datasets . . . . .	102
6.3.2	Implementation Details . . . . .	103
6.3.3	Training Strategies . . . . .	105
6.3.4	Performance of the Proposed Network . . . . .	106
6.3.5	Comparison of SPP and FSPP on PhPAIS Dataset . . . . .	108
6.4	Conclusion . . . . .	109
<b>7</b>	<b>Conclusion and Future Work</b>	<b>110</b>

7.1 Conclusion . . . . . 110

7.2 Future Works . . . . . 111

**Bibliography** **113**

## List of Figures

1.1	Texts are the combination of strokes. . . . .	2
1.2	Stroke-related features used in [111, 30, 120, 27] for text detection and recognition. . . . .	2
1.3	Scene images with horizontal text (selected from ICDAR 2013 dataset [53]) and oriented text (selected from ICDAR 2015 dataset [52]). . . . .	4
1.4	Scene images with multi-lingual multi-oriented text (selected from MLT dataset [91]) and curved text (selected from Total-text dataset [18]). . . . .	5
1.5	Samples of handwritten text. (selected from IAM dataset [86], real bank cheque dataset CVL and CAR [26], and real cell-phone number dataset PhPAIS [128]). . . . .	8
1.6	Structure of LeNet5 [56]). . . . .	12
1.7	Structure of IncepText [141]). . . . .	13
1.8	Structure of RNN and LSTM [2]). . . . .	14
2.1	Structure comparison of VGG, ResNet and Xception. . . . .	39
2.2	The structure of proposed DeepText network (from [126]), where $a@b\#c$ means current block is with kernel size $a \times a$ and output channel $b$ , and is repeated for $c$ times. $S = 2$ means the stride is set to 2 at a specific layer or the last layer of a specific block. . . . .	40

2.3	Atrous convolution, ASPP and depthwise separable convolution (from [13, 14]). . . . .	41
2.4	Output feature maps of proposed network (from [126]). . . . .	43
2.5	Feature extraction by ASPP with the same atrous rates for text with various scales (from [126]). . . . .	44
2.6	Back propagation after using auxiliary losses and connections. The yellow arrows indicate back propagation paths without auxiliary losses and connections, while the green arrows represent additional paths after using auxiliary losses and connections (from [126]). . . .	45
3.1	Overview of one-stage object detector and two-stage object detector.	51
3.2	Network structure of proposed GMask R-CNN. . . . .	53
3.3	Default anchors generated for position $p$ at image scale $l$ and RoIAlign for text proposals. . . . .	55
3.4	Qualitative results of proposed detector. . . . .	59
4.1	Challenging samples of scene text recognition (from [127]). . . . .	61
4.2	Current solutions for scene text recognition (from [127]). When using LSTM, 2-D feature maps are usually converted to 1-D space by pooling or flattening operations. When the LSTM is not used, additional parameters or post-processing steps are involved. . . . .	63

4.3	Overview of proposed FACLSTM (from [127]). $F$ and $M$ denote the extracted feature maps and character centre masks. $T$ groups of feature maps are produced by the proposed attention-equipped ConvLSTM, where $T$ is the maximal string length, and the followed softmax classifier is responsible for producing $T$ groups of feature maps from extracted feature maps. Note that, the softmax classifier and previous fully connected layer are shared by the $T$ groups of feature maps. . . . .	65
4.4	Sampling points in standard convolution and deformable convolution. Blue points are the sampling points and arrows indicate offsets of sampling locations. . . . .	66
4.5	Illustration of deformable convolution. . . . .	67
4.6	Illustration of the FC-LSTM (left) and the ConvLSTM (right) (from [127]). The FC-LSTM is performed in 1-D space, while the ConvLSTM is performed in 2-D space. . . . .	69
4.7	Illustration of our proposed attention-equipped ConvLSTM (from [127]), where the inputs are weighted by attention scores derived from previous cell state and cell output. . . . .	71
4.8	Visualization results of predicted mask and attention shift procedure (from [127]). . . . .	79
4.9	Visualization results of attention predicted by FACLSTM and FLSTM_base1 (from [127]). Values of the attention maps are normalized and truncated for a better visualization. Note that FACLSTM directly produces 2-D attention maps, while FLSTM_base1 generates 1-D attention vectors, which are then reshaped to 2-D space. . . . .	80
5.1	Converting 2D feature maps into 1D space to adapt FC-LSTM to scene text recognition (from [125]). . . . .	82

5.2	The structure of our proposed ReELFA network (from [125]). . . . .	83
5.3	Illustration of the proposed one-hot encoded location (from [125]). . .	84
5.4	End-to-end trainable scene text reading system and related results on licence plate recognition. . . . .	88
6.1	Structure of proposed method (from [128]). . . . .	94
6.2	Resize text images with/without padding. . . . .	95
6.3	Implementation of spatial pyramid pooling (from [128]). . . . .	97
6.4	Feature extraction of the SPP layer and FSPP layer from feature maps with arbitrary sizes (from [128]). . . . .	99
6.5	Connection of the FSPP layer to the next fully-connected layer (from [128]). . . . .	100
6.6	China post mail images used to collect handwritten phone numbers for PhPAIS dataset. Some contents are covered by masks for privacy protection. . . . .	103
6.7	Samples from CVL, ORAND-CAR and PhPAIS. . . . .	104

## List of Tables

1.1	Scene text datasets that widely used in last decade. EN and CN denote English and Chinese, respectively. . . . .	7
1.2	Handwritten text datasets that published in last decade and their recognition performance in ICDAR competitions. Evaluation metrics are word-level accuracy except ones marked by CRA (character level accuracy). . . . .	7
2.1	Comparison with prior arts on ICDAR2015 (from [126]). . . . .	47
3.1	Comparison with state-of-the-art approaches on MLT dataset. Our proposed GMask R-CNN is tested with size 1600 and 1920, indicated by *_1600 and *_1920, respectively. . . . .	58
3.2	Comparison with state-of-the-art approaches on ICDAR2015. . . . .	58
4.1	Result comparison across different methods and datasets (from [127]). Word-level recognition rate is used here. IIIT5K_No, IIIT5K_50 and IIIT5K_1k denote that no lexicon, 50-word lexicon and 1k-word lexicon are used, respectively. Smpls: the number of samples used for training individual models, where * means that datasets derived from SVT are used. . . . .	75



5.1	Results obtained by different methods. ‘IIT5K_*	
	lexicon type used for the evaluation of the IIT5K dataset.	
	‘Ours_noEL’ and ‘Ours_noFA’ represent our model without the	
	encoded location and focused attention respectively. ‘*’ means that	
	the word images containing non-alphanumeric characters are	
	removed from the test dataset. *_bi means binary network setting. .	86
6.1	Distribution of the four different databases with respect to string	
	length . . . . .	104
6.2	Distribution of the cropped datasets with respect to string length .	105
6.3	Recognition accuracies (%) on the cropped datasets . . . . .	106
6.4	Recognition accuracies (%) of different methods . . . . .	107
6.5	Recognition accuracies (%) on the croppedPhPAIS and PhPAIS	
	datasets . . . . .	108

# 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 CF-SPP. CFSPPP 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.