

ARBITRARY-SHAPE SCENE TEXT DETECTION AND ITS APPLICATION IN EDUCATIONAL RESOURCE NAVIGATION

by Chengpei Xu

Thesis submitted in fulfilment of the requirements for
the degree of

Doctor of Philosophy

under the supervision of Prof. Sean He and Dr. Wenjing Jia

University of Technology Sydney
Faculty of Engineering and Information Technology

June 2022

Certificate of Authorship/Originality

I, Chengpei Xu, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology, at the University of Technology Sydney.

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

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program and the UTS FEIT Research Scholarship.

Signature: Production Note:
 Signature removed prior to publication.

Date: June 22, 2022

Acknowledgements

I wish to express my deepest appreciation to my supervisor, Professor Xiangjian He, for his excellent supervision, invaluable advice, and encouragement during my doctoral study. I also wish to express my deepest thanks to Professor He for trusting me and providing me an opportunity to get UTS FEIT Research Scholarship. From the bottom of my heart I know, I can always get help and suggestion from him during my downtime in life and study. It is so true.

I also express my sincere gratitude to my co-supervisor Dr. Wenjing Jia. I still remember how confused I was when I first came to UTS for finding the right research direction. Wenjing's hands-on guidance and insightful analysis helped me to find the right research that I was interested in and good at, making my research journey not boring. I admire Dr. Wenjing Jia's professionalism as a teacher, as I cannot remember how many times Wenjing and I have discussed research issues late into the night.

I also wish to express special thanks to Professor Ruomei Wang, Professor Xiaonan Luo, Professor Zhongxuan Luo, Professor Zhixun Su, Dr. Baoquan Zhao, Dr. Boliang Guan, Dr. Hanhui Li, Dr. Xiaochen Fan, Dr. Qingqing Wang, Mr. Tingcheng Cui, Mr. Lijie Shao, Miss Mengqiu Hu, Mr. Jiachen Kang and Mr. YuanFang Zhang for all their help and valuable discussions.

Most importantly I would like to thank my parents for their years of support. Due to COVID-19, I have not seen them since mid 2019. I have left my hometown Xinjiang for higher education in 2010, four years in Changsha, one year in Guangzhou and seven years in Sydney. I miss them and hope they are proud of what I have achieved today. Finally, I am deeply grateful to my wife Si Wang. I would not have been able to complete my doctorate degree without her dedication.

List of Publications

This thesis is based on the following publications:

- **Chengpei Xu**, Wenjing Jia, Tingcheng Cui, Ruomei Wang, Yuan-fang Zhang, Xiangjian He, Arbitrary-shape Scene Text Detection via Visual-Relational Rectification and Contour Approximation, ‘IEEE Transactions on Multimedia, DOI: 10.1109/TMM.2022.3171085.’ (Accepted; covering most parts of Chapter 3)
- **Chengpei Xu**, Wenjing Jia, Ruomei Wang, Xiaonan Luo, Xiangjian He, MorphText: Deep Morphology Regularized Accurate Arbitrary-shape Scene Text Detection, ‘IEEE Transactions on Multimedia, DOI: 10.1109/TMM.2022.3172547’ (Accepted; covering most part of Chapter 4)
- **Chengpei Xu**, Ruomei Wang, Shujin Lin, Xiaonan Luo, Boquan Zhao, Lijie Shao, Mengqiu Hu, Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos. *in* ‘Proceeding of IEEE International Conference on Multimedia and Expo (ICME), Shanghai, 2019, pp. 898-903’ (covering some parts of Chapter 5)
- **Chengpei Xu**, Wenjing Jia, Si Wang, Ruomei Wang, Xiangjian He, Shujin Lin, Baoquan Zhao, Yuan-fang Zhang, Semantic Navigation of Slide-based Educational Video for AutoNote Generation, ‘IEEE Transactions on Learning Technologies’ (Under review after major revision; covering most parts of Chapter 5)

Contents

Certificate	ii
Acknowledgments	iii
List of Publications	iv
List of Figures	viii
List of Tables	x
Abbreviation	xii
Dedication	1
Abstract	2
1 Introduction	4
1.1 Introduction	4
1.2 Issues and Challenges	7
1.2.1 Arbitrary-shape Scene Text Detection	7
1.2.2 Navigation Tools for Educational Resources	9
1.3 Contributions	10
1.4 Thesis Organization	12
2 Literature Review on Scene Text Detection and Its Application in Navigating Educational Resources	14
2.1 Deep Learning Based Scene Text Detection	14
2.1.1 Different Modeling of Multi-oriented Scene Text Detection	15
2.1.2 Different Modeling of Arbitrary-shape Scene Text Detection	18
2.1.3 Summary of Modeling Methods of Scene Text Detection	19
2.2 Existing Systems and Tools of Navigating Educational Resources	20
3 Arbitrary-shape Scene Text Detection via Visual-Relational Reasoning and Contour Approximation	23
3.1 Introduction	24
3.2 Related Work	30
3.2.1 GCNs based Bottom-up Arbitrary-shape Text Detection	30
3.2.2 False Detection Suppression	31
3.3 Methodology	32
3.3.1 Dense Overlapping Text Segments	33
3.3.2 Graph based Reasoning	38
3.3.3 Visual-Relational Feature Fusion	39
3.3.4 Objective Function	42

3.3.5	Inference by FPNS and Shape-Approximation (SAp)	42
3.4	Experiments	44
3.4.1	Datasets	45
3.4.2	Implementation Details	46
3.4.3	Comparison with State-of-the-art Methods	46
3.4.4	Ablation Studies	50
3.4.5	Limitation	59
3.5	Summary	60
4	MorphText: Deep Morphology Regularized Accurate Arbitrary-shape Scene Text Detection	61
4.1	Introduction	61
4.2	Related Work	67
4.2.1	Removing False Detection in Arbitrary-shape Text Detection	67
4.2.2	Deep Morphological Networks	68
4.3	Methodology	69
4.3.1	Deep Morphological Operations	70
4.3.2	Deep Morphology based Text Segment Regularization	71
4.3.3	Deep Morphology based Relational Reasoning	73
4.3.4	Text Segment Proposal Module	75
4.3.5	Objective Function	76
4.4	Experiments	77
4.4.1	Implementation Details	78
4.4.2	Inference	79
4.4.3	Comparison on Benchmark Datasets	80
4.4.4	Ablation Studies	85
4.5	Limitations	92
4.6	Summary	93
5	Semantic Navigation of Slide-based Educational Video	94
5.1	Introduction	95
5.2	Related Work	98
5.2.1	Navigation and Annotation Tools for Slide-based Educational Videos	98
5.2.2	Presentation Slide Processing	100
5.2.3	Table of Contents and Note Generation	100
5.3	Visual Entity Extraction and Recognition	101
5.3.1	Slide Extraction	102
5.3.2	Visual Entity Extraction	103
5.4	Hierarchical Relationship Generation	105
5.4.1	Visual Saliency of Visual Entities	107
5.4.2	Visual Saliency Clustering	109

5.4.3	Dependency Relation Detection	110
5.5	Semantic Annotation of Visual Entities based on Multi-Channel Information	111
5.6	Annotation Tools and Applications based on Visual Entities for Educational Video	114
5.6.1	AutoNote Generation	114
5.6.2	Table-of-Contents Generation	117
5.7	Experiments	118
5.7.1	Datasets	118
5.7.2	Implementation Details	118
5.7.3	Evaluation of Visual Entity Extraction	120
5.7.4	Evaluation of Hierarchical Relationship Extraction	122
5.7.5	Evaluation of Visual Entity Matching	123
5.7.6	User Study	124
5.7.7	Discussion of Effectiveness in Locating Information	126
5.7.8	Limitation	127
5.8	Summary	128
6	Conclusion and Future Work	129
6.1	Conclusion	129
6.2	Future Work	131
	Bibliography	132

List of Figures

1.1	The illustration of various arbitrary shape text instances in complex environments and discretionary shooting conditions.	5
1.2	Different annotation methods of multi-oriented texts and curved texts	6
1.3	Different modelling examples	7
2.1	The existing navigation systems of educational video	20
3.1	The error accumulation problem of the existing bottom-up approaches (top) and our solution (bottom). Our Graph Guided Text Region fuses relational features with visual features and rectifies false detections. The segment type prediction module further rectifies this through excavating the “characterness” and connectivity of text segments. The Final Output shows that the false detections have been suppressed.	24
3.2	Route-finding gives suboptimal visiting order when there are too many contour points.	28
3.3	The overall structure of the proposed network.	32
3.4	The three types of text segments.	35
3.5	The results of weakly supervised annotation (1st row) and the ground truth (2nd row).	37
3.6	The network structure of multi-modal fusion decoding module.	41
3.7	Visualization of the text detection results obtained on CTW1500 (1st row), Total-Text (2nd row), ICDAR2015 (3rd row) and MSRA-TD500 (4th row).	45
3.8	Some failure cases (1st row) and the ground truth (2nd row).	58
4.1	Both of the GCN-based method [1] and the top-down method [2] have failed (as shown in (a) and (b)) when the text instance is separated due to heavy occlusion. With our morphology regularization, such separated text segments can still be connected into a single text instance (as shown in (c)).	63
4.2	Our proposed MorphText approach effectively addresses the two key issues that restrain the performance of the bottom-up methods. The pink boxes indicate the false detection areas accumulated from the earlier process and the green boxes indicate the disconnected areas.	64
4.3	The overall structure of our network, where “1/4,64”, “1/8,128”,... and “1/32,512” indicate the resize ratio and the channel number.	70
4.4	The network structure of the DMOP module.	71
4.5	The network structure of the DMCL module.	74

4.6	Examples of text detection results obtained with the proposed MorphText approach on benchmark datasets. The first three rows are the results obtained from arbitrary-shape text detection datasets CTW1500 and Total-Text. The last row shows the results obtained from multi-oriented datasets MSRA-TD500 and ICDAR2017-MLT.	79
4.7	Qualitative comparisons with the SOTA methods on challenging samples.	84
4.8	Visualisation of the intermediate results of MorphText addressing noise patterns (top) and the connection problem between text segments (bottom).	91
4.9	Qualitative comparisons with the SOTA bottom-up method [1] (top row) on handling varied sizes of interfering patterns and relative large gaps	91
4.10	Some failure cases of the proposed MorphText, where the green bounding boxes indicate the detected results and the red bounding boxes highlight the failure areas.	92
5.1	The architecture of the proposed annotation pipeline	101
5.2	Visualization of our visual entity extraction. The extracted text, formula and graph entities are enclosed in red, blue and green boxes, respectively.	103
5.3	The overall structure of our visual entity extraction network.	104
5.4	An example of the typical layout of a slide (a) and the hierarchical relationship extracted between its text content blocks (b).	107
5.5	The three classes of English letters	108
5.6	Visual saliency scores for visual entities on a slide	108
5.7	Two different matching situations of corresponding speech texts	113
5.8	A visual example of merging graph entities	115
5.9	Examples of the proposed navigation tool and its applications based on visual entities for educational videos	119
5.10	Time consumption comparison of the Searching task and the Detail Understanding task.	125
5.11	Results of the Summarization Task	127

List of Tables

2.1	The typical top-down and bottom-up methods designed for multi-oriented and arbitrary-shape texts.	16
3.1	Results on CTW1500. (P: Precision, R: Recall, F: F-measure, †: bottom-up methods, §: top-down methods, *: GCN methods)	48
3.2	Results on Total-Text. (†: bottom-up methods, §: top-down methods, *: GCN methods)	49
3.3	Results on ICDAR2015. (†: bottom-up methods, §: top-down methods, *: GCN methods)	50
3.4	Results on MSRA-TD500. (†: bottom-up methods, §: top-down methods, *: GCN methods)	51
3.5	Results on ICDAR2017-MLT. (†: bottom-up methods, §: top-down methods, *: GCN methods)	51
3.6	The impact of our proposed FPNS and SAp strategies.	52
3.7	The impact of the annotation types on the proposed method.	54
3.8	The impact of the width of the text segments on the detection accuracy obtained on CTW1500.	55
3.9	The effectiveness of LAT and FD on CTW1500. (w/o: without) . . .	57
3.10	The effectiveness of the proposed FPNS mechanism on suppressing false positives and false negatives.	57
3.11	Comparison of detection results with different backbones	58
3.12	The time efficiency of the proposed method.	59
4.1	Results on CTW1500. (§: top-down methods, †: bottom-up methods)	81
4.2	Results on Total-Text. (§: top-down methods, †: bottom-up methods)	82
4.3	Results on MSRA-TD500. (§: top-down methods, †: bottom-up methods)	83
4.4	Results on ICDAR2017-MLT. (§: top-down methods, †: bottom-up methods)	83
4.5	The effectiveness of our proposed DMOP and DMCL on CTW1500 and Total-Text.	85
4.6	The impact of the kernel size in DMOP and DMCL on F-measure. The F_{\min} , F_{\max} , F_{mean} and $F_{\text{std-dev}}$ the min, max, mean, and standard deviation for the F-measure.	87
4.7	Comparison of the detection results with/without the residual connection	88
4.8	Comparison of detection results with different NMS thresholds.	89

4.9	Comparison of detection results with DMOP on top-down methods . . .	89
4.10	Inference speeds of the proposed approach on input images of different sizes in different datasets.	90
4.11	The effectiveness of the proposed DMOP/DMCL mechanism on suppressing false positives and false negatives.	92
5.1	Features used to extract the hierarchical relationship	106
5.2	Results of visual entity extraction on ISI-PPT. (*: multi-scale training or testing.)	120
5.3	Results of visual entity extraction on ICDAR2013. (*: multi-scale training or testing)	121
5.4	Comparison of results obtained with different approaches for visual entity extraction.	121
5.5	Performance comparison of hierarchical relationship generation using different features	122
5.6	Performance comparison on headline extraction	123
5.7	Performance comparison on matching corresponding speech text . . .	123

Abbreviation

AAAI AAI Conference on Artificial Intelligence
Adam - Adaptive Moment Estimation
AP - Affinity Propagation
CNN - Convolutional Neural Network
CVPR International Conference on Computer Vision and Pattern Recognition
DMCL - Deep Morphological Closing
DMN - Deep Morphological Network
DMOP - Deep Morphological Opening
ECCV European Conference on Computer Vision
EMD - Earth Mover's Distance
FCN - Fully Convolutional Network
FD - Fuse Decoding
FPN - Feature Pyramid Network
FPNS - False Positive/Negative Suppression
GCN - Graph Convolutional Neural Network
GGTR - Graph Guided Text Region
GPU - Graphical Processing Unit
ICCV International Conference on Computer Vision Recognition
IJCAI The International Joint Conference on Artificial Intelligence
IoU - Intersection over Union
LAT - Location-Aware Transfer
MLT Multi Lingual Text
MM ACM International Conference on Multimedia
MOOC - Massive Open Online Courses
MSRA-TD Microsoft Research Asia Text Detection
NMS - Non-maximum Suppression
OCR - Optical Character Recognition
OER - Open Educational Resources
PR Pattern Recognition
ReLU - Rectified Linear Unit
RNN - Recurrent Neural Network
SAp - Shape Approximation
SE - Structure Element
SGD - Stochastic Gradient Descent
SOTA - State of The Art
TCL - Text Center Line
TIP Transaction on Image Processing
TMM Transaction on Multimedia
TOC - Table of Content
TR - Text Region
WMD - Word Mover's Distance

Dedication

To my wife Si Wang

To my parents Yan Li and Bin Xu

ABSTRACT

ARBITRARY-SHAPE SCENE TEXT DETECTION AND ITS APPLICATION IN EDUCATIONAL RESOURCE NAVIGATION

by

Chengpei Xu

Text instances exist widely as an information carrier in natural scenes, videos and document photos. However, localizing text instances with arbitrary shapes is a challenging task since their style, colour, size, aspect ratio and shape vary greatly depending on the using scenarios. The abovementioned issues hinder the retrieval of information and the digitization of raw photos and videos. The situation worsens when the raw photos and videos are for educational purposes.

In this thesis, we address the challenging problem of arbitrary-shape scene text detection by proposing two deep learning-based bottom-up approaches. Then, we create a navigation system for slide-based educational resources using the semantic information of the detected texts as the primary cue.

In the first approach, we revitalize the GCN-based bottom-up text detection frameworks by aggregating the visual-relational features of text with two effective false positive/negative suppression mechanisms. First, dense overlapping text segments depicting the “characterness” and “streamline” of text are generated for further relational reasoning and weakly supervised segment classification. Then, a Location-Aware Transfer (LAT) module is designed to transfer text’s relational features into visual compatible features with a Fuse Decoding (FD) module to enhance the representation of text regions for the second step suppression. Finally, a novel multiple-text-map-aware contour-approximation strategy is developed, instead of the route-finding process.

In the second approach, targeting building reliable connections between text

segments and alleviating error accumulation in bottom-up modelling, we propose a novel approach to capture the regularity of texts by embedding deep morphology for arbitrary-shape text detection so as to regularize false text segment detection and link missing connections. Towards this end, two deep morphological modules are designed to regularize text segments and determine the linkage between them. First, a Deep Morphological Opening (DMOP) module is constructed to remove false text segment detection accumulated in the feature extraction process. Then, a Deep Morphological Closing (DMCL) module is proposed to allow text instances of various shapes to stretch their morphology in all directions while deriving their connections.

Using the detected arbitrary-shape text information in educational resources as a primary cue, we propose a slide-based video navigation tool that can extract the hierarchical structure and semantic relationship of visual entries in videos by integrating multi-channel information. A clustering approach is proposed for restoring the hierarchical relationship between visual entities. The restored visual entities are then associated with their corresponding audio speech text by evaluating their semantic relationship.

Dissertation directed by Professor Xiangjian (Sean) He

Dissertation co-directed by Dr. Wenjing Jia

Faculty of Engineering and Information Technology

University of Technology Sydney