UTS UNIVERSITY OF TECHNOLOGY SYDNEY

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

List of Figures

List of Tables

Abbreviation

AAAI AAAI 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