CERTIFICATE

It is certified that the work contained in the thesis titled ***"Scene Text Analysis in Unconstrained Environment using Deep Networks"*** by ***Randheer Bagi*** has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all requirements of Comprehensive Examination, Candidacy, and SOTA for the award of Ph.D. Degree.

Supervisor

Dr. Tanima Dutta
Assistant Professor,
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Uttar Pradesh, INDIA 221005.

DECLARATION BY THE CANDIDATE

I, ***Randheer Bagi***, certify that the work embodied in this Ph.D. thesis is my own bonafide work carried out by me under the supervision of ***Dr. Tanima Dutta*** from ***July 2017*** to ***August 2020*** at ***Department of Computer Science and Engineering***, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, *etc.* reported in journals, books, magazines, reports, dissertations, theses, *etc.*, or available at websites and have not included them in this thesis and have not cited as my own work.

Date: 31 August, 2020

Place: Varanasi

(Randheer Bagi)

CERTIFICATE BY THE SUPERVISOR

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(Dr. Tanima Dutta)
Assistant Professor,
Dept. of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi

31.08.2020
Signature of Head of Department
(Prof. Rajeev Srivastava)

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Scene Text Analysis in Unconstrained Environment using Deep Networks

Name of the Student: Randheer Bagi

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the *Doctor of Philosophy.*

Date: 31 August 2020

Place: Varanasi

(Randheer Bagi)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

Dedicated to my parents,

Mrs. Usha Bagi

and

Mr. Satyendra Saroj Bagi

ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisor, Dr. Tanima Dutta, for her invaluable support and assistance. I feel immense pleasure in expressing my profound sense of gratitude and sincere regard for his constant feedback and expertise during all these years. I am eternally grateful to have had the opportunity to work on my thesis under his supervision. I would also like to thank the other members of my Doctoral committee, Dr. Bhaskar Biswas, Department of Computer Science and Engineering, and Dr. Smrity Dwivedi, Department of Electronics Engineering, for their help and support throughout the tenure of my studies. Special thanks to Dr. Hari Prabhat Gupta for his consistent assistance in both work and life aspects. I would also like to convey my sincere gratitude to Dr. Rajeev Srivastava, Head of Department of Computer Science and Engineering and all the RPEC and DPGC members for their suggestions and endorsement to this work.

I am grateful to my colleagues and friends, Surbhi Saraswat, Ashish Gupta, Rahul Mishra, and Aishwarya Soni for the long discussions and their brilliant insights that have helped me to overcome the challenges I have faced in the development of this work.

Finally, I express my heartfelt gratitude to my parents Mrs. Usha Bagi and Mr. Satyendra Saroj Bagi, and my sister Vandana Bagi for their constant support, love, encouragement, and sacrifices. Their affectionate love and care cannot be expressed in words.

(Randheer Bagi)

Contents

List of Figures

List of Tables

List of Symbols

Symbol	Description
$\mathbf{r}$	Region proposal
h, $\mathcal{H}$	Height
w, $\mathcal{W}$	Width
I	Input Image
ζ, C	Number of channels
$\mathbf{f}$	Forget gate
$\mathfrak{W}$, b	Learnable parameters
$\mathcal{P}$, $\mathbf{CP}$	Conditional probability
x	$x-$ Coordinates
y	$y-$ Coordinates
$\circ$	Hadamard product
$*$	Convolution operation
$\sigma(\cdot)$	Sigmoid function
$\mathbf{m}$	Input modulation gate
$\mathbf{i}$	Input gate
$\mathbf{o}$	Output gate
$\mathbf{h}$	Hidden state
α_t	Weight vector
$\mathbf{v}$	Visual features tensor
$SP(\cdot)$	Spatial pooling
$\mathfrak{h}_{t-1}$, $\mathbf{G}_{t-1}$	Previous hidden state
$t-th$	Timestep
RF	Receptive field
$\delta(\cdot)$	Softmax function
$\mathbf{y}$	Character classes

Abbreviations

Abbreviation	**Description**
CTS	Cluttered TextSpotter
CNN	Convolutional Neural Network
ASPP	Atrous spatial pyramid pooling
ROI	Region of- interest
RRT	Ring radius transform
ReLU	Rectified Linear Unit
EOS	End-of-sequence symbol
CDF	Cumulative density function
ITS	Intelligent Transportation Systems
DAS	Driver-Assistance System
DPP	Dilated pyramid pooling
CAA	Context Aggregation Attention
GRU	Gated recurrent unit
LSTM	Long-short term memory
NMS	Non-maximal suppression
CP	Conditional probability
IoU	Intersection-over-union
maxpool	Maxpool Operation
FTS	Fainted TextSpotter
BTS	Blurred TextSpotter
FCN	Fully Convolutional Network
SSD	Single Shot Detector
FPN	Feature Pyramid Network
SFP	Stacked Feature Pooling
GT	Ground Truth

Preface

Scene text analysis aims to detect and recognize text instances from the natural scene images. Scene text detection, recognition, and spotting approaches have received immense attention in the computer vision and multimedia research community. Scene text analysis is widely accepted due to its various real-life applications, such as autonomous vehicles, real-time traffic sign recognition, blind navigation assistance, and multilingual translation. This is however a challenging task since scene text regions have a wide range of scale, orientation, aspect ratio, color, font, language, and script. Such text instances (or regions) can also be in horizontal, oriented, and curved forms. To address the problems of practical interest, we need to consider the presence of partial occlusion, truncation artifact, motion blurs, camera shake artifacts, poor contrast, and faint text edges, which makes the text detection and recognition more complex. In such an unconstrained environment, we have to emphasize attention on salient regions for accurate detection and recognition. Deep network in recent years has gained significant importance in solving computer vision problems, like object detection, scene analysis, text detection and recognition, and image segmentation. However, to the best of our knowledge, the literature that considers the realistic issues, like truncation artifacts, camera shake, and poor contrast is still in the elementary phase.

In this thesis, we address the problem of scene text detection in an unconstrained environment using deep networks. We develop methods that handle the issues like partial occlusion, blurry texts, and faint text edges in scene images. In Chapter 3, we propose a deep learning architecture to address the issue of a cluttered background. The presence of truncated text parts and bizarre artistic style enhances the challenges to analysis. The model focuses attention on local semantics and global structural context of salient features for accurate detection of text instances. In Chapter 4, we develop models for handing blurry scene images. Edges of text instances blurred due to camera shake. We enhance the transformation modeling capability of the salient features and pay attention on pixel-wise spatial information and channel-wise inter-dependencies for precise text localization. Chapter 5 propose a model for addressing faint text edges due to poor contrast. This problem is maneuvered by considering semantic edge supervised feature maps followed by attention on channel-wise relationships of discriminative features. In Chapter 6, we describe an arbitrary-shaped multilingual text spotter that uses deep learning architecture for detection and recognition. It also integrates a learnable non-maximal suppression model to enhance the average precision of the network.

For recognition in all the chapters, we incorporate the Bi-LSTM attention module for recognition of detected text instances and predict script class through majority voting. We also incorporate multi-language character segmentation and word-level recognition in a recognition module. Furthermore, we mitigate the problem of misclassification caused by inter-class interference by exploring inter-class separability and intra-class compactness. We perform a comprehensive set of ablation studies and experiments to show the efficiency of our models. We consider standard metrics, like precision, recall, and f-measure for detection and strong, weak, and generic lexicons for both word spotting and end-to-end recognition. We use publicly available benchmark datasets, such as ICDAR 2013, ICDAR 2015, RRC-MLT 2017, RCTW 2017, CTW1500, Total-Text, MSRA-TD500, COCO-Text, and SVT to demonstrate the efficacy of our text detector network. Our proposed methods outperform the state-of-the-art approaches in terms of recall for detection of text instances in an unconstrained environment.

We develop a new dataset, known as Noisy Arbitrary-shaped Scene Text (NAST) dataset, which contains a large number of noisy scene images that have texts with varying font, scale, orientation, script, and aspect ratio. The images also have partial occlusion, truncated text parts, varying noise density, low contrast, and poor illumination. It also contains images with bizarre artistic styles. It contains 951 images for training and another 321 for testing. The size of images varies from 260×183 to 3436×2700. It has horizontal, multi-oriented, and curve text instances with 7592 text annotations. The ground truth is provided at character, word, and text-line-levels. Most images are collected from Google or captured using phone cameras and also obtained by imposing truncation artifacts on the scene images of publicly available datasets. We also vary the contrast and illumination of benchmark images.

Dataset link: https://drive.google.com/drive/folders/1PdCh2od5lCB- KvP g9PGi3C 2Gehs4reo