

# Contents

<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xvi</b>
<b>List of Symbols</b>	<b>xvii</b>
<b>List of Abbreviations</b>	<b>xix</b>
<b>Preface</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Benchmark Datasets . . . . .	4
1.2 Objectives of the Research Work . . . . .	6
1.3 Contributions of the Thesis . . . . .	8
1.4 Application Scenarios . . . . .	10
1.4.1 Organization of the Thesis . . . . .	11
<b>2 Related work</b>	<b>13</b>
2.1 Scene Text Detection . . . . .	13
2.2 Scene Text Recognition . . . . .	17
2.3 Scene Text Spotting . . . . .	19
2.4 Motivation . . . . .	21
<b>3 Cluttered TextSpotter</b>	<b>23</b>
3.1 Proposed Architecture . . . . .	25
3.1.1 Backbone Network . . . . .	26
3.1.2 Oriented Region Proposal Network . . . . .	28
3.1.3 Context Encoding and Refinement Module . . . . .	30
3.1.4 Inter-class Interference Problem . . . . .	33
3.1.5 Recognition Module . . . . .	34

3.2	Experimental Results . . . . .	38
3.2.1	Ablation Study . . . . .	40
3.2.2	Comparison with State-of-the-Art Results . . . . .	44
3.3	Summary . . . . .	47
<b>4</b>	<b>Blurred TextSpotter</b>	<b>53</b>
4.1	Proposed Architecture . . . . .	55
4.1.1	Backbone Network . . . . .	56
4.1.2	Context Aggregation Attention (CAA) module . . . . .	58
4.1.3	Detection Module . . . . .	61
4.1.4	Recognition Module . . . . .	63
4.2	Experimental Results . . . . .	65
4.2.1	Implementation Details . . . . .	65
4.2.2	Ablation Study . . . . .	67
4.2.3	Comparison with State-of-the-Art Results . . . . .	72
4.3	Summary . . . . .	73
<b>5</b>	<b>Faded TextSpotter</b>	<b>81</b>
5.1	Proposed Architecture . . . . .	83
5.1.1	Semantic Edge Supervised Backbone Network . . . . .	84
5.1.2	Bi-modal Context Encoding . . . . .	85
5.1.3	Localization-aware Oriented Region Proposal Network . . . . .	88
5.1.4	Miss-classification Problem . . . . .	91
5.1.5	Recognition Module . . . . .	91
5.2	Experimental Results . . . . .	94
5.2.1	Implementation Details . . . . .	95
5.2.2	Ablation Study . . . . .	97
5.2.3	Comparison with State-of-the-Art Results . . . . .	102
5.3	Summary . . . . .	104
<b>6</b>	<b>NAST dataset for Multilingual Arbitrary-shaped Scene Text Spotting (MAST)</b>	<b>109</b>
6.1	Proposed Architecture . . . . .	110
6.1.1	Text Mask Computation . . . . .	110
6.1.2	Learnable Polygon Non-Maximum Suppression (LP-NMS) . . . . .	113
6.1.3	multilingual Text Recognition . . . . .	115
6.1.4	NAST Dataset Creation . . . . .	118

---

6.1.5	Autonomous Learning with GT Generation . . . . .	119
6.2	Experimental Results . . . . .	120
6.2.1	Impact of Recognition Module for Word Recognition and Script Identification . . . . .	121
6.2.2	Comparison with the state-of-the-art . . . . .	122
6.2.3	NAST dataset statistics . . . . .	122
6.2.4	Evaluation of Learning using GT Generation . . . . .	123
6.3	Summary . . . . .	124
<b>7</b>	<b>Conclusion and Future Work</b>	<b>129</b>
	<b>References</b>	<b>132</b>
	<b>List of Publications</b>	<b>146</b>



# List of Figures

1.1	Exemplification of Problem 1. . . . .	7
1.2	Exemplification of Problem 2. . . . .	8
1.3	Exemplification of Problem 3. . . . .	8
1.4	Exemplification of Problem 4. . . . .	9
3.1	Illustrating the necessity of Cluttered TextSpotter in scene text spotting. Columns (b) and (c) are the recognized text instances in the scene images of column (a) using baseline [1] and our network. Cluttered TextSpotter can spot oriented text instances in the scene images with a cluttered environment. . . . .	24
3.2	The architecture of the proposed Cluttered TextSpotter.. . . . .	25
3.3	Architecture of the backbone network. . . . .	28
3.4	Architecture of the oriented region proposal network. . . . .	29
3.5	Architecture of the context encoding and refinement module. . . . .	31
3.6	Architecture of the recognition module. . . . .	35
3.7	Effect of datasets on power consumption for different devices. . . . .	46
4.1	Illustration of the necessity of Blurred TextSpotter. Columns (b) and (c) are the recognized text instances in the scene images of column (a) using baseline [1] and our network. It can spot oriented text instances in the blurry scene images. . . . .	54
4.2	Overall Architecture of Proposed Network. . . . .	55
4.3	Architecture of backbone network. . . . .	57
4.4	Architecture of context aggregation attention module. . . . .	59
4.5	Architecture of recognition module. . . . .	63
4.6	Effect of datasets on power consumption for different devices. . . . .	72

5.1	Illustration of natural images (first row) representing the necessity of Fainted TextSpotter. Second and third rows are the recognized scene text instances using FOTS [1] (baseline) and the proposed network, respectively.	82
5.2	Overall architecture of Fainted TextSpotter. . . . .	84
5.3	Architecture of the backbone network. . . . .	84
5.4	Architecture of bi-modal context encoding module. . . . .	86
5.5	Illustration of Fused Dilated Convolutions. . . . .	87
5.6	(a) The line chart about the counting number of components in different orientations. (b) The red line correspond to center of the line chart. . .	89
5.7	Architecture of the proposed recognition module. . . . .	92
5.8	Effect of devices on power consumption for different datasets. . . . .	99
6.1	Illustration of natural images (first row) representing the necessity of our multilingual text spotter. Second row is the recognized scene text instances by the proposed network. . . . .	110
6.2	Illustration of natural images (first row) representing the necessity of our network for detecting curve text instances. Second row is the recognized scene text instances by the proposed network. . . . .	111
6.3	Overall architecture of arbitrary shaped text spotter. . . . .	111
6.4	Illustration of IoU values of an arbitrary-shaped text mask. . . . .	112
6.5	Illustration of text mask prediction in mask branch. . . . .	113
6.6	The architecture of learnable polygon maximal suppression model. Here, $(i, j, z)$ in a mask represents the resolution $(i, j)$ and depth $(z)$ of the feature map produced in that mask. . . . .	114
6.7	Architecture of the proposed recognition module. . . . .	115
6.8	Statistics of NAST dataset. . . . .	127

# List of Tables

3.1	Effect of different variations of backbone network over ICDAR 2015 dataset.	41
3.2	Effect of variation in dilation rate on ICDAR 2015 [2] and NAST dataset.	41
3.3	Effect of different branches of context encoding and refinement module.	42
3.4	Effect of different softmax functions on COCO-Text dataset. . . . .	42
3.5	Effect of variation in size of RoI in detection on ICDAR 2013 [3] and NAST dataset. . . . .	43
3.6	Effect of variation in scale of RoI in detection on ICDAR 2013 [3] and NAST dataset. . . . .	43
3.7	Effect of different branches of recognition module on ICDAR 2015 and NAST dataset. . . . .	44
3.8	Effect of variation in the number of channel in text spotting on ICDAR 2015 [2] dataset. . . . .	44
3.9	Effect of variation in size of RoI in text spotting on ICDAR 2015 [2] dataset. . . . .	45
3.10	Effect of variation in scale of RoI in text spotting on ICDAR 2015 [2] dataset. . . . .	45
3.11	Specifications of the smartphones with Adreno-640 GPU that are used for experimentation. . . . .	46
3.12	Performance comparison on SVT dataset. . . . .	47
3.13	Performance comparison on MSRA-TD500 dataset. . . . .	48
3.14	Performance comparison on COCO-Text dataset. . . . .	49
3.15	Performance comparison on ICDAR 2013 [3] and ICDAR 2015 [2] dataset.	50
3.16	Performance comparison on SVT dataset. . . . .	51
3.17	Performance comparison on ICDAR 2013 dataset for the recognition. .	51
3.18	Performance comparison on ICDAR 2015 dataset for the recognition. .	52

3.19	Test time speed in terms of FLOPS, number of training parameters, and frames per second (FPS) on ICDAR 2015 dataset for detection (D), recognition (R), or spotting (S). . . . .	52
4.1	Impact of different variations of MobileNetV2, IGCV2, and ShuffleNetV2 as backbone networks on ICDAR 2015 dataset. . . . .	68
4.2	Performance comparison on ICDAR 2015 [2] and NAST dataset, where M stands for <b>maxpool2D</b> , $2 \times 2$ , stride = 1. . . . .	69
4.3	Effect of variation in size of RoI in detection on ICDAR 2013 [3] and NAST dataset. . . . .	69
4.4	Effect of variation in scale of RoI in detection on ICDAR 2013 [3] and NAST dataset. . . . .	70
4.5	Impact of different branches of attention module. . . . .	70
4.6	Effect of different branches of recognition module over COCO-Text and NAST dataset. . . . .	70
4.7	Effect of variation in the number of channel in text recognition on ICDAR 2015 [2] dataset. . . . .	71
4.8	Effect of variation in size of RoI in text spotting on ICDAR 2013 [3] dataset. . . . .	71
4.9	Effect of variation in scale of RoI in text spotting on ICDAR 2013 [3] dataset. . . . .	72
4.10	Specifications of the smartphones with Adreno-640 GPU that are used for experimentation. . . . .	73
4.11	Performance comparison on ICDAR 2013 [3] dataset for text detection in scene images. . . . .	74
4.12	Performance comparison on ICDAR 2015 [2] dataset for text detection in scene images. . . . .	75
4.13	Performance comparison on MSRA-TD500 dataset. . . . .	76
4.14	Performance comparison on COCO-Text [4] dataset for text detection in scene images. . . . .	76
4.15	Performance comparison on SVT [5] dataset for text detection in scene images. . . . .	77
4.16	Performance comparison on ICDAR 2013 datasets for the recognition. . . . .	77
4.17	Performance comparison on ICDAR 2015 datasets for the recognition. . . . .	78
4.18	Performance comparison on COCO-Text [4] and SVT [5] dataset for text recognition in scene images. . . . .	78



---

4.19	Test time speed in terms of on FLOPS, number of training parameters, and frames per second (FPS) on ICDAR 2015 dataset for detection (D), recognition (R), or spotting (S). . . . .	79
5.1	Effect of different variations of MobileNetV2, ShuffleNetV2 and IGCV2 as backbone networks on ICDAR 2015 dataset. . . . .	97
5.2	Effect of context encoding on COCO-Text and SVT datasets. . . . .	98
5.3	Effect of different softmax functions on COCO-Text Dataset. . . . .	98
5.4	Specifications of the smartphones with Adreno-640 GPU that are used for experimentation. . . . .	99
5.5	Effect of different modules of recognition branch over COCO-Text and NAST dataset. . . . .	100
5.6	Effect of variation in size of RoI in detection on ICDAR 2013 [3] and NAST dataset. . . . .	100
5.7	Effect of variation in scale of RoI in detection on ICDAR 2013 [3] and NAST dataset. . . . .	100
5.8	Effect of variation in the number of channel in text recognition on ICDAR 2015 [2] dataset. . . . .	101
5.9	Effect of variation in size of RoI in text spotting on ICDAR 2015 [2] dataset. . . . .	101
5.10	Effect of variation in scale of RoI in text spotting on ICDAR 2015 [2] dataset. . . . .	101
5.11	Performance comparison on ICDAR 2013 [3] dataset for text detection in scene images. . . . .	102
5.12	Performance comparison on ICDAR 2015 [2] dataset for text detection in scene images. . . . .	103
5.13	Performance comparison on MSRA-TD500 dataset. . . . .	104
5.14	Performance comparison on COCO-Text dataset for detection of texts in scene images. . . . .	105
5.15	Performance comparison on SVT dataset for detection of texts in scene images. . . . .	105
5.16	Performance comparison on ICDAR 2015 dataset for word spotting and end-to-end recognition of texts in scene images. . . . .	106
5.17	Performance comparison on COCO-Text and SVT datasets for text recognition accuracy and word spotting in scene images. . . . .	106
5.18	Test time speed in terms of on FLOPS, number of training parameters, and frames-per-second (fps) on ICDAR 2015 dataset. . . . .	107

---

6.1	Effect of different modules of recognition branch over RRC-MLT 2017 and NAST datasets for word recognition. . . . .	122
6.2	Effect of different modules of recognition branch over RRC-MLT 2017 and NAST datasets for script identification. . . . .	122
6.3	Curve text detection performance on Total-Text [6] dataset. . . . .	123
6.4	Performance comparison on RRC-MLT 2017 dataset for detection of text in scene images. . . . .	124
6.5	Performance comparison on RCTW 2017 datasets for detection of text in scene images. . . . .	125
6.6	Performance comparison on RRC-MLT 2017 dataset for text recognition and script identification in scene images. . . . .	125
6.7	Performance comparison on SCUT-CTW1500 [7] for curve text detection in scene images. . . . .	126
6.8	Curve text recognition performance on Total-Text [6] dataset. . . . .	126
6.9	Performance on Our Dataset with GT Learning. . . . .	126

# List of Symbols

Symbol	Description
$\mathbf{r}$	Region proposal
$h, \mathcal{H}$	Height
$w, \mathcal{W}$	Width
$I$	Input Image
$\zeta, C$	Number of channels
$\mathbf{f}$	Forget gate
$\mathfrak{W}, \mathbf{b}$	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
$\alpha_t$	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

<b>Abbreviation</b>	<b>Description</b>
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
<b>IoU</b>	Intersection-over-union
<b>maxpool</b>	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