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