Chapter 1

Introduction

A text is a combination of edges, connected components, strokes, and textures that give meaningful information to the viewer and reader. Text instances in scene image contain high-level important semantic information, which will help in analyzing and understanding the scenes. Texts contain many important clues, such as notices, door-plates, captions, and traffic signs that helps in scene understanding. Due to the rapid advancement of technology such as resource-constrained devices, like smartphones, mobile computing device, etc., which enhance the active research field in computer vision. These devices help in capturing a huge amount of scene images around the world. Captured scene images contain the lots of meaningful information which is needed to extract for gaining the information. Therefore, text plays an important role in the analysis of scene image. It is easy to detect the text from the printed document rather than scene images due to the large variations in backgrounds, textures, fonts, and illumination conditions. Scene images found great attention due to its various practical applications, such as the robot and drone navigation, driver assistance system, text preserving animations, and human-computer interaction. Thus, the scene text analysis can help in real-time applications. Scene text analysis is broadly categories into three ways, like scene text detection, scene text recognition, and scene text spotting. Text detection

and recognition are complement to each other, whereas spotting is combination of both.

- **Scene Text Detection**: It is a process of identifying the presence of text in a scene image and localize the text instances using bounding boxes or polygons.
- **Scene Text Recognition**: It is defined as the assigning the label to the detected text instances using prior knowledge.
- **Scene Text Spotting**: It is the process of detecting and recognizing text instances in scene images in a unified manner.

Text plays a dominant role in analyzing and understanding as text carries rich and important information relevant to the contents. Moreover, some studies have shown that humans often pay more attention to text over other objects in an image because text helps in getting the semantics relevant to the content of the images. To extract relevant information from scene image text detection and recognition methods are used. This method in scene images plays a vital role in the field of computer vision, augmented reality and other innovations. It helps in removing noise in scene images and identify semantic text. Scene text, spotting is a process of simultaneous localization and recognition of all text instances, signs, and markings in a natural scene image. In the current scenario, it is challenging due to the large variations in text shape, orientations, aspect ratios, font size, font style, script, background variation, and symbols. Although traditional methods handle these challenges in the scene images. There are additional challenges, *i.e.,* the occurrence of noise text instances due to fainted text edges, camera motion (motion blur), camera shake, partial occlusions, and cluttered background.

A crucial challenge in scene text analysis is to handle the noisy edges, especially in complex scene images acquired in an unconstrained environment. In most scene text analysis methods, this problem is not considered in their subsequent methods. Detecting and recognizing faint and blue edges or partial occlusion artifacts in scene images is important because edges make shapes and boundaries and provide an opportunity for detecting a meaningful text. Noisy scene images are common in the various domains

due to scene images captured in limited visibility. For example, low contrast scene images or with a cluttered background in which the shape of the text may be noisy due to shading effects, images captured in poor lightning, or less time of exposure. Noisy edges imposed a significant problem due to variability along edges during capturing a scene image in an unconstrained environment. In this scenario, text, spotting becomes a crucial issue.

Traditional methods handle the challenges, like wide variation in text size, orientation, aspect ratio, color,script, and font in the scene images. However, many years of research, there will be set of grand challenges may till be encountered during detection and recognition of text in the scene images. Scene text analysis in an unconstrained environment is still a challenging problem. In an unconstrained environment, addresses the issue of noises, like partial occlusion, distortion, truncation, and inter-class interference. The integration of deep neural network (in short deep networks) with the vision-based problems re-formulate the traditional scene text detection and recognition problem in a new paradigm to provide better performance.

We highlight the challenges that exist in the state-of-the-art methods for application as follows:

Challenge 1: Data unbalanced, low intraclass compactness and inter-class separability. The huge diversity in scene text types, such as road marking, traffic sign symbols, and license plate numbers. Scene text has different scales, font, aspect-ratio, language, and orientation, which adversely affects the efficiency of spotting.

Challenge 2: Diverse characteristics. The scenic images may be captured by the moving vehicles, where texts in such low-resolution images are far away from the mounted camera. Texts appear small, noisy, and blurry, making the process of spotting more challenging in real-time processing.

Challenge 3: Class imbalance. As per our knowledge, the study on the uni-

fied network for handling detection and recognition of scene texts, traffic signs, road-marking, and car license plate to meet the real-time requirement of practical applications is still in the elementary phase.

CHALLENGE 4: TRADE-OFF. It is a complicated task to balance a trade-off between the accuracy and speed in an end-to-end trainable network in terms of inference time. The fine-tuning of accuracy and speed may mitigate human error, which in turn will facilitate precise spotting in day-to-day life applications.

1.1 Benchmark Datasets

The publicly available benchmark datasets that are used for pretrain the models and for experimentation are summarized as follows:

• **Synthtext dataset** [8] is a popular synthetic dataset for scene text detection and recognition. It has a huge number of multi-oriented text instances that are annotated with character-level and word-level rotated bounding boxes. It is composed of 800k images having 10 synthetic words per image, which are placed on real scene background. An image is annotated with a ground truth word and not at a character-level. Each text instance is annotated with its text-string, word-level, and character-level bounding-boxes.

• **ImageNet dataset** [9] is a large-scale image database. It is quite accurate and diverse in nature. It contains 80000 noun synsets of WordNet and has 500-1000 clean and full-resolution images to illustrate each synset. ImageNet has tens of millions of annotated images that are built upon by the semantic hierarchy of WordNet [10]. ImageNet is an image dataset organized according to the WordNet hierarchy. It has on average 1000 images to illustrate each synset.

ICDAR 2013 dataset [3] is a dataset that focuses on the detection and recognition of horizontal text instances in natural scene images. There are 255 images in the training set with 716 annotated words and 233 images in the test set. Apart from the

bounding box, transcriptions are also assigned for each character-level and word-level text instance.

ICDAR 2015 Incidental Text dataset [2] contains 1000 training and 500 testing images, collected using Google glasses and of somewhat low resolutions. It has multi-oriented text instances in each image, having word-level annotations for bounding boxes.

SVT dataset [5] is composed of images from Google Street View that consist of frontal texts of street names, pavement markings, and shop names. SVT set uniquely suited for word spotting. It has 647 images of cropped words with lower resolutions and perspective distortion. Also, the dataset has only word-level annotations (no character bounding boxes). We use a lexicon containing 50 words for each image, known as SVT-50.

COCO-Text dataset [4] is one of the largest and challenging dataset that composed of 43686 training images and 20000 testing or validation images. It has text instances in arbitrary orientations. Some images in this dataset have blur edges. It has more than 63000 images and more than 145000 text instances. Text instances categorized into machine printed and handwritten text. Text instances categorized into English script and non-English script.

MSRA-TD500 dataset [11] contains 500 indoor and outdoor scenes images that are captured using a pocket camera. The indoor office and mall images contain caution plates, signs, and door plates, while the outdoor street images are mostly guided billboards and boards in English and Chinese languages with complex backgrounds. The resolutions of the images vary from 1296×864 to 1920×1280. The training set contains 300 images randomly selected from the original dataset and the remaining 200 images constitute the test set. All the images in this dataset are fully annotated.

RRC-MLT 2017 dataset [12] consists of multi-language and multi-oriented text instances in scene images. It is widely applicable for the task like multi-lingual text detection, crop word script identification, and joint text detection and script identifi-

cation. It contains 18000 images, which comprised of the text of six scripts belonging to nine languages, *i.e.*, Arabic, Bangla, Chinese, English, French, German, Italian, Japanese, and Korean. The training set has a total of 9000 images for nine languages. It has 2000 images per language, however, an image could contain the text of more than one language.

Total-Text dataset [6] is a dataset for curve text detection and recognition. It features curved-oriented texts. It has with 4265 curved texts out of 9330 total text instances. Total-Text is split into 1255 training and 300 testing images. It has conventional horizontal and multi-oriented text along with the curved-oriented text. Total-Text is highly diversified in orientations, more than half of its images have a combination of more than two orientations.

SCUT-CTW1500 dataset [13] is a curve text dataset, which includes over 10000 text annotations in 1500 images. It uses a 14-point polygon-based curved text detector to detect curved text without using an empirical combination. It contains both English and Chinese text instances. It has 1000 images as a training set and 500 as a testing set.

RCTW 2017 dataset [14] a very large Chinese text dataset. It has 1 million Chinese characters from 3850 unique ones annotated by experts in over 30000 street view images. This is a challenging dataset with good diversity containing planar text, raised text, the text under poor illumination, distant text, and partially occluded text. Most images are collected in the wild by phone cameras. Some are screenshots. The images exhibit various kinds of scenes, including street views, posters, menus, indoor scenes, and screenshots of phone apps.

1.2 Objectives of the Research Work

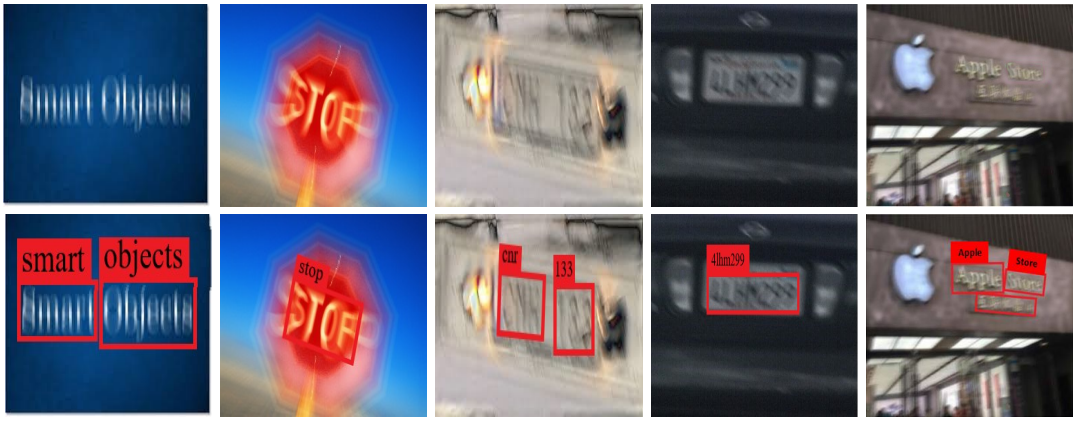
The objectives of this thesis are based on the challenges that arise in capturing scene images in everyday life. This address different problems that frequently arise in real-

world applications. We briefly describe the problems as follows:

- PROBLEM 1: In practical applications, scene images have many objects scattered in the background. Text instances are smaller in nature. The presence of background clutters, partial occlusion, and truncation of texts are a typical issue in scene images. For accurate spotting of scene texts, it is important to overcome the background clutters and partial occlusion artifacts, as shown in Figure 1.1. The presence of cluttered background noise restricts in dense feature matching process and leads to misclassification problem.

Figure 1.1: Exemplification of Problem 1.

- PROBLEM 2: Nowadays, scene images are captured very often using mobile phones, cameras, and webcams. The issue of camera shake and motion blur are commonly found while taking images unless images are captured by experts. The perspective distortion and blurry/shaky text edges make the process of spotting very complex, as shown in Figure 1.2.
- PROBLEM 3: In day-to-day applications, the effect of weather conditions in terms of ambient lighting and noises impacts a lot while capturing a scene image. The presence of adverse weather conditions and ambient noises, such as fog, rain, poor contrast, and low illumination are the main reasons for the fainted text edges in the scene images, as shown in Figure 1.3. This enhances the problem of inter-class interference in classification.



(a) (b) (c)

Figure 1.2: Exemplification of Problem 2.

Figure 1.3: Exemplification of Problem 3.

- PROBLEM 4: Texts in scene images are printed with artistic style in billboards, shop name, door plates, and so on. Different scripts and languages are used across the world and within the countries. The detection and recognition tasks of arbitrary-shaped multi-lingual text instances are important for images captured in practical applications, as shown in Figure 1.4.

1.3 Contributions of the Thesis

In this thesis, we investigate the deep network architecture models for scene spotting in an unconstrained environment. We develop robust scene text spotters that can detect

Figure 1.4: Exemplification of Problem 4.

and recognize text instances with high accuracy. We focus on the practical challenges that occur while capturing scene images in real-life scenarios. The main contributions addressed by the thesis are summarized as follows:

- SOLUTION 1: We humans when we have to identify a text region in a scene image, no matter how complicated the background clutters are, the localization of text instance is subserved by both a local process aware of position-sensitive features and a global process of retrieving structural context.

  We follow a similar approach to solve the issue of background clutters. We incorporate a multi-attention mechanism that learns both the local part information and the global structural context from the semantically enriched feature maps to successfully overcome the issue of partial occlusion.

- SOLUTION 2: In case of blurry and shaky scene images, we human try to understand the context related to text instances and then pay attention on transformation invariant features to overcome the effect of perceptual distortion.

  We model our network based on the aforementioned process by capturing multi-scale context information followed by enhancing the transformation modeling capability of network to classify scene texts in presence of camera shake, motion blur, and geometric distortions.

- Solution 3: While identifying a text instance in a scene image with fainted edges, we humans first emphasis on all the edges in the image (both fainted and non-fainted) and then focus attention of edges which seems to associate with the text instances.

  We pursue the same model by learning semantic edge supervised feature maps followed by focusing attention on pixel-wise spatial features and channel-wise relationships. This helps to resolve the problem faint text edges due to poor contrast.
- Solution 4: We have designed a 10-point polygon based mask for detecting arbitrary-shaped text instances and trained the recognition module multi-lingual annotated datasets for recognition multi-language text instances.

1.4 Application Scenarios

Scene text detection and recognition is widely used in various practical applications, which are used in day-to-day life. A brief description of various applications are as follows:

1. Autonomous Driving / Driving Assistance System : While on the move, The driving assistance system captures scene images around the roadside. It utilizes the meaningful information of such images to ensure safe driving.
2. Blind Navigation: Assisting the visually impaired along their navigation path is a challenging task which drew the attention. It can use meaningful information written in images for better clarity.
3. Multilingual Translation: Across the world, within the countries, different languages and scripts are used. It is important to understand that only a native translater of the target language can translate into it. Automated systems can enhance the performance.
4. Scene Understanding: While analyzing scenes, text information present in

scene images plays an important role in understanding the scene.

5. Text Preserving Stylization: While stylizing the images and videos, due to the use of operators, the text information sometime may lost. Text preserving stylization help in animating images and videos while persisting the text information.

1.4.1 Organization of the Thesis

The rest of thesis is organized as follows.

1. The next chapter elaborates the state-of-the-art literature for scene text detection, recognition, and spotting using deep networks. We also describe the motivation of the thesis based on the drawbacks of recent literature.
2. Chapter 3 proposes a text spotter that can address the issue of cluttered environment of scene images. It is an end-to-end trainable deep neural network that uses local part information, global structural features, and context cue information of oriented region proposals for spotting text instances. It helps to localize in scene images with background clutters, where partially occluded text parts, truncation artifacts, and perspective distortions are present. We mitigate the problem of misclassification caused by inter-class interference by exploring inter-class separability and intra-class compactness.
3. Chapter 4 proposes a robust text spotter for text spotting in blurry scene images. We address different noises, like a motion blur, Gaussian blur, camera shake noise, and inter-class interference. We apply a multi-scale contextual information enriched encoder-decoder based backbone network followed by a spatial and channel-wise attentions. An oriented region proposal network is used for obtaining text proposals. A Bi-LSTM with GRU attention mechanism is incorporate for text recognition.
4. Chapter 5 proposes an end-to-end trainable deep neural network that can ad-

dress the issue of spotting oriented text instances in scene images, captured in adverse meteorological conditions. It localizes words and performs word spotting for every rotated bounding box. It is a scene text spotter that utilizes hierarchical spatial context, channel-wise inter-dependencies, and semantic edge supervision to localize and recognize words in scene images. We explore inter-class interference to reduce the misclassification problem. A recognition module for character segmentation and word-level recognition is incorporated using Bi-LSTM and self attention mechanism.

5. Chapter 6 describes a process for obtaining arbitrary-shaped text masks, multilingual text recognition, and predicts script class for the instances in the scene images. It also incorporate learnable non-maximal suppression mechanism.
6. In the last chapter, we conclude the thesis and summarize the future research direction of the work presented in the thesis.