# CHAPTER 1

# INTRODUCTION

This chapter includes the following subtopics, namely: (1) Rationale; (2) Statement of the problem; (3) Hypothesis (Optional); (4) Theoretical Framework; (5) Conceptual Framework/Paradigm; (6) Scope and Delimitation; and (7) Importance of the study.

## 1.1 Rationale

Scene text detection is a system that aims to find and localize text regions from a natural scene images. Scene text detection has been developing in recent years. It is due to its numerous practical applications in textual image understanding. An effective scene text detection can improve the performance of several multimedia applications such as automatic language translation, content-based image retrieval, robot navigation, autonomous driving, assisting visually impaired people, etc.

Detecting text in natural scene images are highly challenging and more complex than text in documents. The challenges of scene text detection can be roughly categorized into three aspects [40]: (1) Diversity of text in natural scenes such as, different fonts, sizes, colors, shapes, orientations, scales, languages, etc; (2) Complexity of background. For example, elements with extremely similar patterns with text (e.g., bricks, fences, leaves, trees, and traffic signs) or occluded text caused by another objects that may cause confusions and errors; (3) Imperfect imaging quality. For example, low contrast, blur, and non-uniform illumination. Numerous scene text detection methods have been proposed and summarized in [20, 36]. There are several inspiring methods that have been proposed to detect text in natural scene images. These methods can be roughly categorized into three mainly groups: (1) Connected component (CC) based methods, which use component analysis (e.g., stable region, color clustering, stroke width) to extract character candidates; (2) Sliding window based methods, which use a classifier to search for every possible text regions by sliding windows; (3) Deep learning based methods, which use end-to-end feature learning.

Recently, most methods of scene text detection are built upon deep learning models [20]. The methods have obtained promising results, and able to deal with various challenging scenarios like long text, multi-oriented text, multilingual text, and multi-scale text. However, the performance of deep learning model or machine learning model mainly depends on the training data. The model can achieve a perfect prediction when the input is the same or similar to the data which its trained on, yet output can be false prediction (either false positive or false negative) when the input is completely new which was never trained on [5].

## 1.2 Statement of the Problem

As stated in the previous section, the deep learning based methods may output false positive prediction when the input is completely new which was never trained on. Another approach for detecting text in natural scene images (without learning) is connected component based methods. The most popular method for this approach are Maximally Stable Extremal Regions (MSER) [25] and Stroke Width transform (SWT) [4]. These methods have been able to deal with several challenging scenarios. However, these methods are sensitive to low contrast and small text.

## 1.3 Objective

The objective of this thesis is to design and develop a scene text detection method by combining the deep learning and traditional method. The traditional method is expected to improve the deep learning performance by extracting text candidates. The text candidates extraction method should be robust to small text and low contrast. A contrast enhancement method should be exploited to enhance the image contrast. A candidate text regions extraction method should be investigated to improve the detection result on f-score.

## 1.4 Hypotheses

Features that discriminate text and non-text are: (1) a text has nearly constant width, (2) a text has nearly similar color, (3) a text has invariant and stable property. The traditional method (connected component and sliding window based methods) usually refer to these definitions to find the text candidates. To tackle the small text problem, the second definition might be useful. Since the text has nearly similar color, despite the text size variation a text can be distinguished from the background by its color. The text regions can be found by dividing the image into homogeneous regions with quadtree.

For the low contrast problem, a contrast enhancement technique can help to improve the image contrast. Histogram equalization (HE) is a method for adjusting image intensities using histogram. Applying HE on each RGB color channels is not a proper way, because the equalization involves the image intensity and not the color components. Therefore, a color conversion from RGB to L*a*b* is more suitable. Finally, HE is performed on the L* channel (lightness) and converted back to RGB for further processes.

The advantage of deep learning based methods is automatic feature learning from the input images. A convolutional neural network (CNN) model could be adopted for detecting text in natural scene. A candidate text regions extraction (CTR) method based on traditional methods can be used to filter the result of CNN. The CTR divides the image into homogeneous regions with quadtree. The result of the CTR is a CTR map. The CTR map is expected to be able to filter the false positive prediction.

## 1.5 Theoretical Framework

Traditional (non-learning) based methods typically explore handcrafted features like color, edge/gradient, texture, and stroke features for detecting text in natural scene. Color features are exploited by the assumption that text has a consistent and and the color is distinguishable against its background [16]. Edge features are exploited by the assumption that text has a strong gradient against its background [16]. Texture features are exploited by the assumption that text consist of dense character, which is distinguishable against its background [36]. Stroke features are exploited by the assumption that text has a consistent stroke width [4]. As stated in the previous section, the most popular methods for this approach is MSER and SWT. MSER assumes that the color of each character within a word is consistent, and SWT assume the stroke width are consistent. As stated in the problem statement, these methods are sensitive to small text and low contrast.

Deep learning based methods typically consist of end-to-end trainable model. The model automatically learns the features on the training data. Most of deep learning models are data-hungry. Their performance would be accurate when sufficient data are available. There are various approaches to train the model, and can be roughly categorized as detection and segmentation based approach. Recent detection based approach consists of 2 steps end-to-end trainable model: predicting the existence of text, and regressing the text location [20].

Contrast enhancement is a technique to improve the image contrast. The typical methods to perform contrast enhancement are histogram equalization and histogram specification. Histogram equalization adjust the image intensities using the image histogram, and histogram specification transform the image histogram to match a specified histogram. The histogram equalization method is more suitable for scene text image, because the adjustment may increase or decrease the contrast, so that an extremely low or high contrast regions is adjusted to nearly normal contrast.

Image segmentation is a process to divide the image into multiple segments. Based on the assumption that a text has nearly constant color and different color against the background, scene text image can be segmented to divide the image into homogeneous regions and extract the candidate text regions. To achieve this, a quadtree technique can be used. A quadtree is obtained by recursively dividing the image into four quadrants. The color homogeneity with a standard deviation value is useful for determining whether the region should be divided or not. Finally, the candidate text regions should be the group of regions in the quadtree leaves. Based on this process, despite the text size variation, text regions should be able to be extracted.

## 1.6 Conceptual Framework/Paradigm

The existing methods for detecting text in natural scenes have their own advantages and disadvantages. The deep learning based methods allow the automatic feature learning. However, their performance highly depend on the availability of data. This study proposes a method to assist the model by providing a candidates of text with traditional methods, thus may possibly reduce the false positive prediction without re-train the model. The schematic diagram of the scene text detector is shown in Figure 1.1.

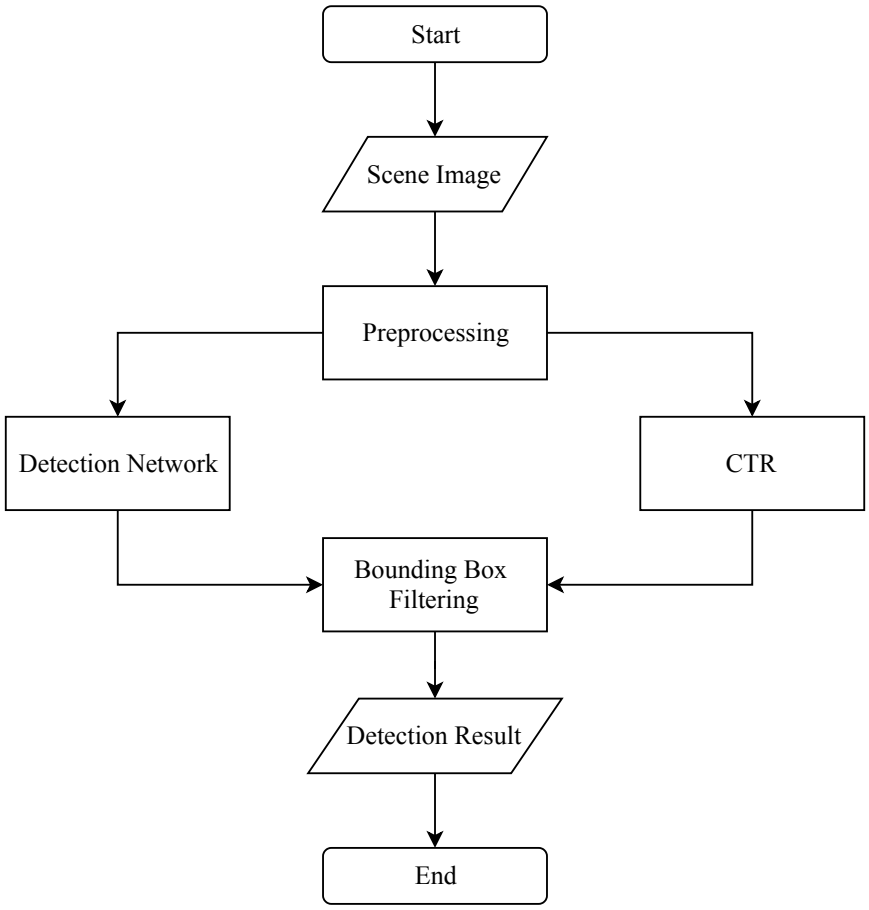


Figure 1.1: Schematic diagram of the scene text detection

## 1.7 Scope and Delimitation

In this research, the scope and delimitation of this study are:

1. The datasets used throughout this study are ICDAR 2013, ICDAR 2015, MSRA-TD500.

2. The output of proposed scene text detector is a quadrilateral bounding boxes that indicate the text location.

## 1.8 Significance of the Study

The proposed method exploits color features by evaluating homogeneous areas with quadtree and extracting the text candidates with neighbour color similarity. The results of this study can contribute to the following usage: for small text, low contrast, multi-oriented, multi-lingual, and multi-scale problems. Furthermore, the proposed preprocessing and CTR can improve the detection performance.