Text Localization in Natural Scene Images using Extreme Learning Machine

A thesis submitted in partial fulfilment of the requirement for the

Degree of Master of Computer Application

of

Jadavpur University

By

Sudeshna Konar

Registration No. : 133689 of 2015-16

Examination Roll No. : MCA186025

Under the Guidance of

Prof. Mahantapas Kundu

Department of Computer Science and Engineering

Jadavpur University, Kolkata- 700032

India

May, 2018

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled **"Text Localization in Natural Scene Images using Extreme Learning Machine"** has been satisfactorily completed by Sudeshna Konar (University Registration No. : 133689 of 2015-16, Examination Roll No. : MCA186025). It is a bonafide piece of work carried out under my guidance and supervision and be accepted in partial fulfilment of the requirement for the Degree of Master of Computer Application, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, Kolkata.

_____

Prof. Mahantapas Kundu (Thesis Supervisor)
Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Countersigned

_____

Prof. Ujjwal Maulik
Head, Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032

_____

Prof. Chiranjib Bhattacharjee
Dean, Faculty of Engineering and Technology
Jadavpur University, Kolkata-700032

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled **"Text Localization in Natural Scene Images using Extreme Learning Machine"** is a bonafide record of work carried out by Sudeshna Konar in partial fulfilment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University during the period of January 2018 to May 2018. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

_____

Signature of Examiner

Date:

_____

Signature of Supervisor

Date:

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis entitled **"Text Localization in Natural Scene Images using Extreme Learning Machine"** contains literature survey and original research work by the undersigned candidate, as part of her Degree of Master of Computer Application.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name : Sudeshna Konar
University Registration No. : 133689 of 2015-16
Examination Roll No. : MCA186025

Thesis Title: Text Localization in Natural Scene Images using Extreme Learning Machine

Signature:

_____

Date:

_____

ACKNOWLEDGEMENT

First and foremost, I would like to start by thanking God Almighty for showering me with the strength, knowledge and potential to embark on this wonderful journey and to persevere and complete the embodied research work satisfactorily.

I am pleased to express my deepest gratitude to my supervisor, **Prof. Mahantapas Kundu**, Department of Computer Science and Engineering, Jadavpur University, Kolkata for his invaluable guidance, constant encouragement and inspiration during the period of my dissertation. I am also very thankful to **Dr. Nibaran Das**, Department of Computer Science and Engineering, Jadavpur University, Kolkata for his guidance and help.

I am highly indebted to **Jadavpur University** for providing me the opportunity and the required infrastructure to carry on my thesis.

I am also grateful to the **Center for Microprocessor Applications for Training Education and Research** for giving me the proper laboratory facilities as and when required.

I would also like to acknowledge all my lab mates especially **Ms. Kalpita Dutta** for helping me and motivating me constantly and spending such wonderful six months journey.

I am thankful to all the teaching and non-teaching staff whose helping hands have smoothed my journey through the period of my research.

Last but not the least, I would like to thank my family members, classmates, seniors and friends for giving me constant encouragement and mental support throughout my work.

Sudeshna Konar
University Registration No. : 133689 of 2015-16
Examination Roll No. : MCA186025
Master of Computer Application
Department of Computer Science and Engineering
Jadavpur University

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

In recent years text localization and recognition from images have been increasingly popular because text images represent many technical and digital information which are used in the different fields of computer vision. Text, as the physical form of language, is one of the basic tools for preserving and communicating information. Text is one of the principal medium of communications which can be found in scattered form throughout the images and which is available in different fonts, colors and shapes. Text data present in images contain useful information for automatic annotation, indexing and structuring of images. Extraction of this information involves detection, localization, tracking, extraction, enhancement, and recognition of the text from a given image. Methods for scene text localization and recognition find all the areas in an image that would be considered as text by a human, mark boundaries of the text areas and output a sequence of characters associated with its content. They are used to process images taken by a digital camera or a mobile phone and to read the content of each text area into a digital format. So, in today's world visual detection and recognition of text from image is very claimable because of its application in content-based image retrieval, robotic navigation, automatic car number plate recognition, extracting information from passport or business card or bank statement, making editable the text of an image, text translation on mobile phones etc.

Broadly speaking, scene text analysis includes locating and identifying text contents from natural scene images. So, text localization and recognition tend to be quite challenging due to the uncontrolled image capturing process, variations of font style, size, color, orientation, contrast, context, geometric and photometric distortion of text in scenes etc. Additionally, text-like background objects, such as bricks, windows and leaves, often lead to many false positive in text detection. The major challenges [7] can be roughly categorized into three types:

DIVERSITY OF SCENE TEXT:

Instead of regular font, single color, consistent size of texts, natural scene images normally contains entirely different fonts, colors, scales of texts.

COMPLEXITY IN THE BACKGROUND:

The backgrounds of the images can be very complex. Elements like bricks, grasses, storefronts, street signs and different types of signs are difficult to distinguish from the true texts and thus confusion or error may occur.

INTERFERENCE FACTORS:

Various interference factors like noise, blur, low resolution, distortion may cause failure in the text detection phase.

Many of these difficulties are found in images from standard text localization and recognition datasets and fail to deal with or ignoring any of these issues will cause the accuracies or detection rates of the text localization and recognition algorithm that is being developed to significantly decrease. Dealing with all these difficulties is a trivial task but to tackle all of these challenges a rich body of approaches have been proposed and substantial progresses have been achieved in recent years. In scene text detection and recognition, representation involves the way and manner of describing and modelling text and background in natural scene images.

1.1 Brief Overview of Text Localization and Recognition

Methods for scene text localization and recognition detect all the sub regions in an image that would be considered as text by a human. In general, this text detection can be done in the following ways,

1. Text detection by locating text in bounding boxes [1][2].
2. Text extraction by binarizing the scene images such that all text pixels are foreground and the rests are background [3][4].
3. Text region proposals methods giving multiple possible text bounding boxes [5][6].

Category 1 and 2 relate to explicit location of text and third category depends on a recognition module to correctly locate the text. In the last two decades, several approaches to text detection have been pursued. Most of the works covered in this thesis were presented after 2010 and the majority of them deal with either text localization or end-to-end text localization in natural scene images.

1.2 Objective of the Work

The objective of the work is to improve the state-of-the-art text localization task on the benchmark dataset ICDAR-2015. To do that, we have used MSER (Maximally Stable Extremal Region) and a classifier ELM [8] (Extreme Learning Machine) specialized in text localization using HOG (Histogram of Oriented Gradients) feature. HOG-based shape detector has been proved to be effective in text detection. In addition, HOG-based methods have false positives mainly from objects and complex background with shape close to text contents. We observed that a large portion of the false positives come from pole structured objects, building outliers and storefronts. Extreme learning machine (ELM) is proposed as an efficient single-hidden-layer

neural network which performs quite well for pattern recognition task with fast learning speed. It has been shown as an effective learning method in a wide variety of applications. So, in this thesis, a detection method based on HOG and ELM is proposed in order to improve the result.

1.3 THESIS OUTLINE

This thesis has been organized into five main chapters. The first chapter presents a rough idea about the domain of text localization and recognition, different challenges and the motivation behind the current work. The second chapter highlights the evolution of different text detection methods over the past few years and provides a literature survey of the novel approaches utilized in the avenue of text localization and recognition. Also a brief introduction of the benchmark datasets has also been included in this section. The third chapter provides a detailed study of a scene text detection method Maximally Stable Extremal Region (MSER) using Extreme Learning Machine (ELM) classifier exploring HOG feature. Here, the proposed methodology has been mathematically established for the localization of the texts. The fourth chapter validates the efficiency of the method in localizing texts in natural scene images. Here the corresponding challenges related to our work and the advantages of using this method is also described. The last chapter concludes and discusses relevant future work in scene text understanding.

CHAPTER TWO

LITERATURE REVIEW

In this chapter some fundamental concepts and related works that are useful to understand the methods on text localization and recognition are introduced and briefly described.

2.1 TEXT LOCALIZATION AND RECOGNITION METHODS

2.1.1 REGION-BASED

The main objective of segmentation is to divide an image into multiple regions. Some segmentation methods such as thresholding achieve this goal by looking for the boundaries between regions based on discontinuities in grayscale or color properties. Region-based segmentation is a technique for determining the region directly. These region-based algorithms can be classified into two main classes:

MERGING ALGORITHMS:

In merging algorithms neighboring regions are compared and if they are close enough in some property then they are merged.

SPLITTING ALGORITHMS:

In splitting algorithms large regions which are not uniform in nature, they are broken up into smaller areas in order to make them uniform.

There are algorithms which are also a combination of splitting and merging. In each of this cases, some criterion is applied to decide whether the regions should be merged or split into multiple regions. This criterion is defined by using different statistical approaches such as standard deviation, variance on the measured property (intensity, color, mean etc.) of the regions.

Among them the most relevant region-based methods proposed in literature throughout the last decade, the two works that are worth mentioning due to their particular novelty are of Pan et al. [9] and Wang et al. [10].

Pan et al. [9] have proposed a novel hybrid method to accurately localize texts in natural scene images. They have built a text confidence map with the help of a text region detector based on which text components can be segmented by local binarization method. A Conditional Random Field (CRF) model is then proposed to label the components as “text” or “non-text”. This proposed method gives promising performance comparing with the existing ones on ICDAR 2003 competition dataset.

The second algorithm which is worth analyzing is of Wang et al. [10] where two systems are developed to solve the end-to-end problem of word recognition. The first one represents a two-stage pipeline consisting of text detection followed by Optical Character Recognition (OCR) engine. The second one is a system of generic object recognition.

But nowadays, region-based methods have become less popular after introduction of connected component-based approaches which are able to overcome all the limitations of region-based methods such as: (i) high computational complexity, (ii) long training times and (iii) false-positive detection errors as many regions of natural scene images are difficult to distinguish from the text components.

2.1.2 CONNECTED COMPONENT-BASED

The main objective behind connected component-based approach is that usually text characters have the same geometric properties. For example, the color of a text character remains the same for the whole letter, the characters are usually placed on a high contrast background to increase readability. So, researchers have been done over the years to propose different techniques on text localization/recognition.

2.1.2.1 MAXIMALLY STABLE EXTREMAL REGION (MSER)-BASED

Most connected component-based text localization and recognition methods propose Maximally Stable Extremal Region (MSER) [11] to identify the text components from a natural scene image.

MSER is a method for blob detection [12] in images to compute a number of co-variant regions from a given gray image called MSER. A MSER is a stable connected component of some gray-level sets of the image that stays nearly the same through a wide range of thresholds. In text detection, due to the geometric properties of text characters, MSER have become quite popular during the last few years.

For this very reason many researchers tried to increase the rates (the percentage of ground-truth text elements identified as stable components) of MSER method by proposing different algorithms. For example:

Neumann et al. [1] presents an end-to-end real-time scene text localization and recognition method. This method leads to a coverage rate of 95% over ground-truth character for ICDAR 2011 dataset. This is an outstanding coverage result for ICDAR 2011 but the amount of extracted ER is very high as compared to MSER. Thus the algorithm requires a constant time O(1) for every ER to maintain computational complexity.

Multi-channel MSER [13] is another efficient variant of the MSER algorithm. This paper is the first to report both text detection and recognition results on the standard and challenging dataset ICDAR 2003. The performance of the method [13] is also evaluated on the another dataset Chars74k on which a recognition rate of 72% has been achieved which is 18% higher than the previous ones.

Tian et al. [12] proposes a multi-level MSER technology that identifies the best-quality text candidates from a set of stable regions that are extracted from different color channel images. The proposed method is evaluated on the ICDAR2003 and SVT datasets and experiments show that it outperforms both popular document image binarization methods and state of the art scene text segmentation methods.

Nafla et al. [14] presents an efficient method of scene text detection using two machine learning classifiers: one for generating candidate word regions and the other for the classification of text or non-text components. Then with the help of a support vector machine classifier they classify a block into text and non-text components.

2.1.2.2 STROKE WIDTH TRANSFORM (SWT)-BASED

Stroke Width Transform (SWT) [15] is another algorithm that is quite popular among text localization works. Unlike MSER-based algorithms SWT does not use threshold values for an image at multiple levels to find the stable connected components. Instead this method looks for edges in the images (using Canny Edge Detector algorithm [16]) and builds a stroke width map in which the value assigned to each pixel denotes the width of the edge it belongs to. Here are some works worth mentioning:

This paper [15] presents a novel image operator to find the value of stroke width for each pixel of an image and uses it for the task of text detection in natural images. The operator is local and independent of data which makes it fast and eliminates the need for multi-scale computation. This algorithm is able to detect texts of different fonts and languages because of its simplicity.

Huang et al. [17] presents a new approach for text localization by dividing text and non-text regions at three levels. Firstly they propose Stroke Feature Transformation (SFT) which extends to SWT and secondly based on the SFT results they uses two classifiers: a text-component classifier and a text-line classifier. It is evaluated on two benchmark datasets ICDAR 2005 and ICDAR 2011 and the corresponding f-measure values are 0.72 and 0.73 respectively.

Risnumawan et al. [18] proposes a novel method to obtain stroke width image without the edge detectors by replacing the edge detector algorithm with the extremal regions (ERs) and presenting a novel weighted Markov Random Field (MRF) method with three properties to construct a finer stroke width image. Experiment results on ICDAR datasets and a comparison with the state-of-the-art methods have shown the efficiency of the proposed method.

2.1.3 DEEP-BASED

Convolutional Neural Networks (CNN) [19] have recently been successfully used for text localization and recognition in natural scene images. Since deep-based methods are very recent, there are few deep-based text spotting works in literature. In this section, the three most relevant works that have been proposed during the last few years are discussed in a summarized manner.

Coates et al. [20] realize the power of features learned by a CNN at identifying text characters from natural images. However, they cannot find a proper way to identify bounding boxes for text elements. Therefore they do not participate in the ICDAR challenge to evaluate how their features perform when evaluate using ICDAR's evaluation protocol. Instead, they evaluate the precision/recall of their deep-based model using per-pixel based accuracies computed from the ground-truth annotations for ICDAR 2003 images.

In [21] a single very large CNN has been used for integrated text localization and recognition of Street View House Numbers [20] (SVHN) and CAPTCHA, thus removing the need of using local windows or proposals as in region-based and CC-based methods.

In [22] multiple very large CNN trained solely on synthetic data are used to localize and read text word proposals from Edge Box and ACF detector.

2.2 BRIEF DESCRIPTION OF SOME POPULAR DATASETS

There are several benchmark datasets [23] related to text localization and recognition which are as follows-

2.2.1 CHARS74K DATASET

CHARS74K dataset [40] was collected by de Campos et al [24]. It contains 7705 English character images (A-Z, a-z and 0-9, in total 64 classes) and 3345 Kannada character images (647 classes), which were manually segmented from 1922 scene text images. Additionally, the dataset contains 1416 scene images with each word annotated by a polygon and its text transcription, however not every word in the dataset is annotated.

2.2.2 ICDAR-2003 DATASET

The dataset was created by Simon Lucas and his colleagues for the ICDAR-2003 Robust Reading Competition [25]. The dataset was used in an unchanged form in the ICDAR-2005 Robust Reading Competition, which is why in the literature sometimes the dataset is also referred to as the ICDAR-2005 dataset. It contains 258 training and 251 testing images with words and characters annotated by bounding boxes and their text content. 1157 word and 6185 character images (1111 word and 5430 character images) were subsequently cropped from the training (respectively testing) image set, to be used in Cropped Word Recognition and Cropped Character Recognition

evaluation. The dataset was captured by people who were specifically tasked to capture text in an outdoor environment, so as a result text in the dataset is mostly horizontal, it occupies a large portion of an image and it typically is present in the middle of an image since the authors of the pictures tried to capture “nice” pictures of text.

2.2.3 ICDAR-2011 DATASET

It [41] was created by taking all images from the ICDAR-2003 dataset, removing images with no text, adding several new images and splitting them again into a training and a testing subset. The dataset was first used in the ICDAR-2011 Robust Reading Competition and then subsequently in the 2013 Competition, which is why it is sometimes referred to as the ICDAR-2013 dataset. In the ICDAR-2015 Robust Reading Competition [26], the dataset was used again in the Focused Scene Text Challenge. ICDAR-2011 contains 229 training and 255 testing images, with corresponding 849 training and 716 testing cropped word images. As a result of the creation process, the testing subset of the ICDAR-2011 dataset contains the same images as the training subset of the ICDAR-2003 dataset. This unfortunately often leads to evaluation problems in the literature, where some methods are trained on both ICDAR-2003 and 2011 training sets, falsely assuming they are different datasets, and evaluated on the 2011 testing set but the testing set contains many images from the joint training set, and therefore the accuracy evaluation is heavily affected.

2.2.4 ICDAR-2015 DATASET

It was collected by people wearing Google Glass devices [39] and walking in Singapore and then subsequently by selecting and annotating only images with text. The ICDAR-2015 dataset was introduced in the ICDAR-2015 Robust Reading Competition [40] to address the problems of the ICDAR 2003/2011 datasets. The dataset is used in the Incidental Scene Text Challenge. The dataset contains 1670

images with 17548 annotated words. 1500 images are publicly available, split into training and testing set, and the remaining 170 images represent a sequestered set for a future use. Each word is annotated by a quadrilateral (3 points) and its Unicode transcription, thus supporting rotated and slanted text. The images in the dataset were taken “not having text in mind" and therefore contain a high variability of text fonts and sizes and they include many realistic effects (e.g. occlusion, perspective distortion, blur or noise etc.).

2.2.5 STREET VIEW TEXT (SVT) DATASET

It [42] was published by Wang and Belongie, where the data was collected by asking annotators to find images with local businesses in the Google Street View application. The dataset contains mostly business names and business signs and the business names by looking up businesses close to the GPS position of the image. In total, the dataset contains 350 images (100 training and 250 testing images) of 20 different cities and 725 labeled words. The word annotations were also exploited to create the dataset of cropped words *SVT-50*, which contains 647 word images, each with a lexicon of 50 words. There is also a lexicon of all test words (4282 words), which is referred to as *SVT-FULL*. The annotators were instructed to find a representative text associated with the business in the image, then to move the viewpoint in the application to minimize the skew of the text and finally to save the screenshot. The words in the image picked by the annotators in this process are tagged by a horizontal bounding-box and a case-insensitive transcription.

Apart from the above mentioned datasets there are few more types of datasets related to text localization and recognition. They are [23] (a) COCO-TEXT DATASET, (b) IIIT DATASETS, (c) KAIST DATASET, (d) NEOCR DATASET. The details of the above mentioned datasets are given in the following table:

Table 2.1: Comparative Study on Different Benchmark Datasets

Dataset Name	***Collected By***	***Description***
CHARS74K [40]	Campos et al. [24]	1922 scene text images, 1416 scene images and 7705 English and 3345 Kannada character images which were manually segmented from 1922 scene text images
ICDAR-2003 [25]	Simon Lucas and his colleagues [25]	258 training and 251 testing images and 1157 word and 6185 character images were cropped from the training image set
ICDAR-2011 [40]	Created from ICDAR-2003 removing images with no text and adding several new images	229 training and 255 testing images and 849 training and 716 testing cropped word images
ICDAR-2015 [39]	Google Glass Devices [39]	1670 images with 17548 annotated words and 1500 images are publicly available
SVT [42]	Wang and Belongie	100 training and 250 testing images of 20 different cities and 725 labeled words
COCO-TEXT [44]	Based on the MSCOCO dataset	63 686 images and 173 589 annotated words
KAIST [45]	Hyung et al. [45]	3 000 images of indoor and outdoor scenes with text
IIIT [46]	Harvested from Google image search	2000 training and 3000 testing images
NEOCR [47]	Nagy et al. [47]	659 real world images with 5238 text line annotations

CHAPTER THREE

METHODOLOGY

We have presented our work on one of the popular text related dataset of scene images ICDAR-2015 which consists of 229 training images and 233 testing images. Here, in this chapter we are trying to describe the whole proposed methodology through the following steps-

- Maximally Stable Extremal Region (MSER) detection
- Removal of non-text regions based on Stroke Width Variation
- Merging of text regions for the final detection result
- Feature extraction exploring Histogram of Oriented Gradients (HOG)
- Classification using Extreme Learning Machine (ELM)

3.1 MAXIMALLY STABLE EXTREMAL REGION (MSER) DETECTION

3.1.1 BRIEF INTRODUCTION OF MAXIMALLY STABLE EXTREMAL REGION (MSER)

MSER [23] [24] is a popularly used method for blob detection in images. The MSER algorithm extracts a number of co-variant regions from an image. An MSER is a stable connected component of some gray-level sets of the image. MSER is based on the idea of taking regions which stay nearly the same through wide range of thresholds. All the pixels below a given threshold are white and all those above or equal are black.

If we are shown a sequence of thresholded images I_t with frame t corresponding to threshold t, we would see first a black image, then whitespots corresponding to local intensity minima will appear then grow larger. These whitespots will eventually merge, until the whole image is white. The set of all connected components in the sequence is the set of all extremal regions. The word extremal refers to the property that all pixels inside the MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary.

3.1.1.1 MATHEMATICAL DETAILS

Let us consider an image I(x), $x \in \Lambda$ is a real function of a finite set Λ with a topology τ. Elements of Λ are called pixels.

For simplicity, let us take $\Lambda = [1, 2, \ldots, N]^n$ and the topology τ induced by the 4-way or 8-way neighborhoods, but we do not restrict ourselves to n = 2 as [27].

A level set S(x), $x \in \Lambda$ of the image I(x) is the set of pixels that have intensity not greater than I(x), i.e.

$$S(x) = \{y \in \Lambda : I(y) \leq I(x)\} \qquad \text{.....Equation 1}$$

A path $(x_1, \ldots, x_n)$ is a continuous sequence of pixels (i.e. such that x_i and x_{i+1}) are 4-way or 8-way neighbors for $(i = 1, \ldots, n-1)$. A connected component C of the set Λ is a subset $C \subset \Lambda$ for which each pair $(x_1, x_2) \in C^2$ of pixels is connected by a path fully contained in C.

The connected component is maximal if any other connected component C′ containing C is equal to C. An extremal region R is a maximal connected component of a level set S(x). We denote by R(I) the set of all extremal regions of image I.

STABILITY CRITERIA-

Among all extremal regions R(I), we are interested in the ones that satisfy certain stability criteria which will be introduced next. Let the level I(R) of the extremal region R be the maximum image value attained in the region R, i.e.

$$I(R) = \sup_{x \in R} I(x) \qquad \text{.....Equation 2}$$

Now, an extremal region R of a one dimensional image I(x) is shown and the corresponding extremal regions are considered as $R_{+\triangle}$ and $R_{-\triangle}$. Stability is computed based on the area variation of such regions.

Let $\triangle > 0$. Let $R_{+\triangle}$ be the smallest extremal region that contains iR and has intensity which exceeds of at least $\triangle$ the intensity of R, i.e.

$$R_{+\triangle} = argmin\{|Q|: Q \in R(I), Q \supset R, I(Q) \geq I(R) + \triangle\} \qquad \text{.....Equation 3}$$

Similarly, let $R_{-\triangle}$ be the biggest extremal region containing R that has intensity which is exceeded by at least $\triangle$ by R, i.e.

$$R_{-\triangle} = argmax\{|Q|: Q \in R(I), Q \subset R, I(Q) \leq I(R) - \triangle\} \qquad \text{..... Equation 4}$$

Consider the area variation,

$$\rho(R;\triangle) = \frac{|R_{+\triangle}| - |R_{-\triangle}|}{|R|} \qquad \text{..... Equation 5}$$

The region R is maximally stable [28] if it is a minimum for the area variation, in the following sense: $\rho(R;\triangle)$ is smaller than $\rho(Q;\triangle)$ for any extremal region Q "immediately contained" or "immediately containing" R.

It is said that an extremal region R immediately contains another extremal region Q if $R \supset Q$ and if R' is another extremal region with $R \supset R' \supset Q$, then $R' = R$. Note that this notion makes sense because the base set Λ is finite.

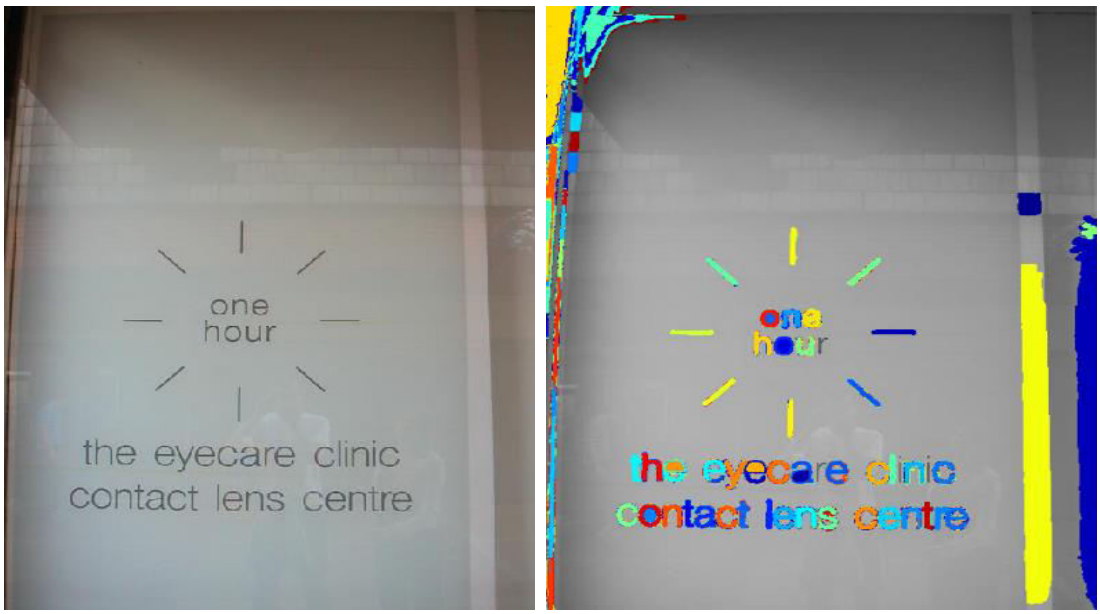
3.1.2 Implementation Of MSER In Our Work

First, most of the candidate text regions are detected using MSER. It works well for text regions because the high consistent color and high contrast of text leads to stable intensity profiles.

In MSER method, the default 'RegionAreaRange' is 30 to 14,000 and the default 'ThresholdDelta' is 2. But, we have used the range as 80 to 8000 and the threshold value as 4 after doing an analysis of text images in the ICDAR-2015 dataset. 'RegionAreaRange' indicates the size of the region in pixels. The two-element vector [minArea, maxArea], allows the section of regions containing pixels to be in between minArea and maxArea, inclusive.

'ThresholdDelta' indicates the step size between intensity threshold levels and a numeric value in the range (0,100]. This value is expressed as a percentage of the input data type range used in selecting extremal regions while testing for their stability. Decrease of this value returns more regions.

So, in our work, we have experimented with different values of ThresholdDelta. As it decreases the more regions (less portion of text and a large portion of non-text) are detected as MSER. So, in order to detect the candidate text regions properly (comparatively less portion of non-text) we have increased the ThresholdDelta value and decreased the variation between minArea and maxArea.



(a) (b)

Figure 1: (a) A sample input image from ICDAR-2015 dataset and (b) exhibits all the potential text regions detected using MSER

3.2 Removal of Non-Text Regions Based on Stroke Width Variation

3.2.1 Brief Introduction of Stroke Width Variation

Stroke Width Variation [29] is a common metric used to discriminate between text and non-text. Stroke width is a measure of the width of the curves and lines that make up a character. Text regions tend to have little stroke width variation, whereas non-text regions tend to have larger variations. To help understand how the stroke width can be used to remove non-text regions, estimate the stroke width of one of the detected MSER regions. You can do this by using a distance transform and binary thinning operation. In the images shown below, there is how the stroke width

image has very little variation over most of the region. This indicates that the region is more likely to be a text region because the lines and curves that make up the region all have similar widths, which is a common characteristic of human-readable text.

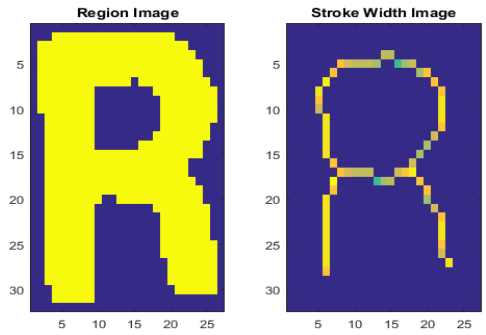


Figure 2: Variation [25] of stroke width image over the region of the left side

3.2.2 Implementation Of Stroke Width Variation In Our Work

In the Figure 1(b), as we can see there are many non-text regions detected along with the text components. So, we have removed the non-text regions based on Stroke Width Variation.

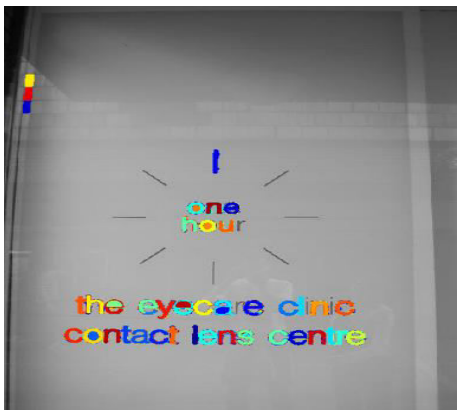


Figure 3: After removal of non-text regions based on Stroke Width Transform

3.3 MERGING OF TEXT REGIONS FOR THE FINAL DETECTION RESULT

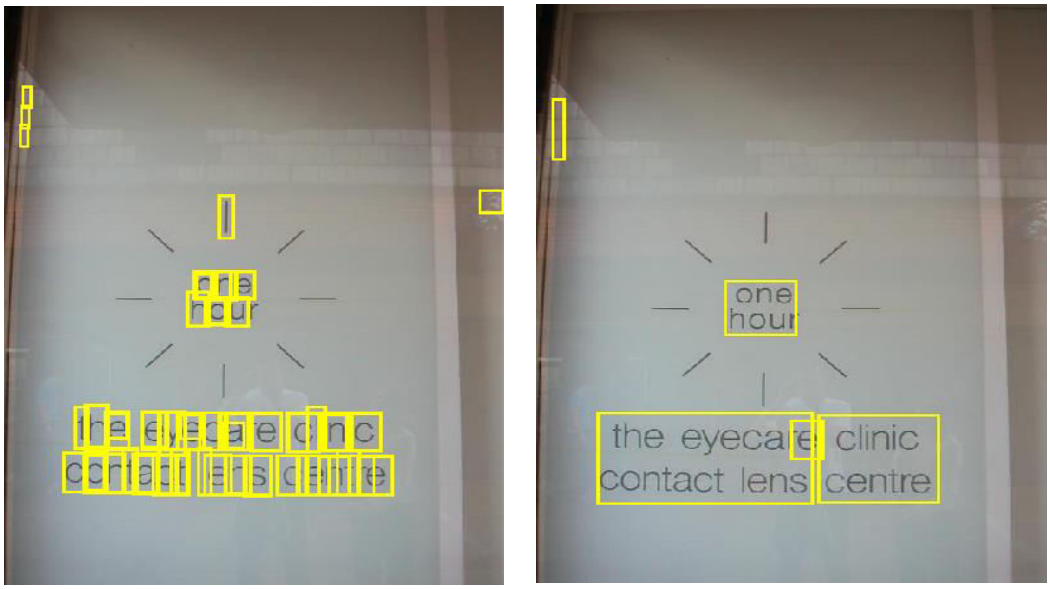
At this point, all the detection results are composed of individual text characters. To use these results for further tasks, the individual text characters must be merged into words or text lines. This enables detection of the actual words in an image, which carry more meaningful information than just the individual characters.

For example [25], recognizing the string 'EXIT' vs. the set of individual characters {'X','E','T','I'}, where the meaning of the word is lost without the correct ordering.

One approach for merging individual text regions into words or text lines is to first find neighboring text regions and then form a bounding box around these regions. To find neighboring regions, the bounding boxes computed earlier with regionprops will be expanded. This makes the bounding boxes of neighboring text regions overlap such that text regions that are part of the same word or text line form a chain of overlapping. Now, the overlapping bounding boxes can be merged together to form a single bounding box around individual words or text lines.

To do this, we have computed the overlap ratio of all bounding box pairs. This denotes the distance between all pairs of text regions so that it is possible to find groups of neighboring text regions by looking for non-zero overlap ratios.

Once the pair-wise overlap ratios are computed, all the text regions connected by a non-zero overlap ratio are achieved.



(a) **(b)**

Figure 4: (a) Individual text characters and (b) detected merged word or text lines

3.4 Feature Extraction Exploring Histogram of Oriented Gradients (HOG)

3.4.1 Brief Introduction Of Histogram Of Oriented Gradients (HOG)

In recent years, HOG (Histogram of Oriented Gradients) algorithm [30] has got popularity in different applications. Researchers tend to use HOG algorithm for recognizing objects in images. HOG algorithm is used object recognition with very high success rate.

3.4.1.1 Mathematical Details

HOG [31] [32] is a gradient-based feature set which is computed densely on the image. First, the input image is normalized. Then the x-derivatives and y-derivatives

of the image are computed. From x and y-derivatives the gradient magnitude and orientation for a pixel (x, y) is computed as follows:

$$\theta(x,y) = atan2\left(G_x(x,y), G_y(x,y)\right) \quad \text{.....Equation 6}$$

$$m(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \quad \text{.....Equation 7}$$

Here, m(x, y) is the gradient magnitude, θ(x, y) is the gradient orientation, $\mathbf{G_x}(\mathbf{x},\mathbf{y})$ and $\mathbf{G_y}(\mathbf{x},\mathbf{y})$ are the x and y-derivatives at pixel-position (x, y). The gradient responses are accumulated over local cells into local gradient orientation histograms. The orientations are quantized into discrete bins modulo 180 degrees. Each pixel votes with its gradient magnitude for the entry in the histogram corresponding to its gradient orientation. Cells are then further grouped into overlapping blocks. To achieve robustness each block is then normalized. The final feature vector is built by concatenating all normalized blocks in the detection window.

In our work after merging text regions we have extracted feature exploring HOG. We have used HOG feature for the reasons:

(i) capture edge or gradient structure in HOG is very characteristic of local shape,

(ii) it is relatively invariant to local geometric and photometric transformations,

(iii) within cell rotations and translations do not affect the HOG values,

(iv) illumination invariance achieved through normalization,

(v) the orientation sampling densities can be tuned for different applications (such as human detection, hand gesture detection, text detection etc.

3.5 CLASSIFICATION USING EXTREME LEARNING MACHINE (ELM)

3.5.1 BRIEF INTRODUCTION OF EXTREME LEARNING MACHINE (ELM)

Extreme learning machine (ELM) [8] is proposed as an efficient single-hidden-layer feedforward neural network. It has been shown as an effective learning method in a wide variety of applications. The efficiency of ELM provides the possibility to incorporate part-based model into naturalistic driving data process which normally has high requirement for speed.

Huang et al. [33] shows that dual optimization objective functions of ELM is consistent with that of SVM (Support Vector Machine) while ELM searches optimal solution in a greater domain with faster implementation. Therefore ELM achieves better performance in general and multiple tests have also proved it. So, in our work a detection method based on HOG and ELM is proposed in order to improve the processing speed.

3.5.2 MATHEMATICAL DETAILS

Huang et al. [33] theoretically and experimentally proved that ELM can be used as a unified learning platform which does not need to tune the hidden layer parameters as traditional neural networks do. Instead of using the time-consuming gradient descent based learning method; ELM relies on computing the inverse of the hidden layer matrix.

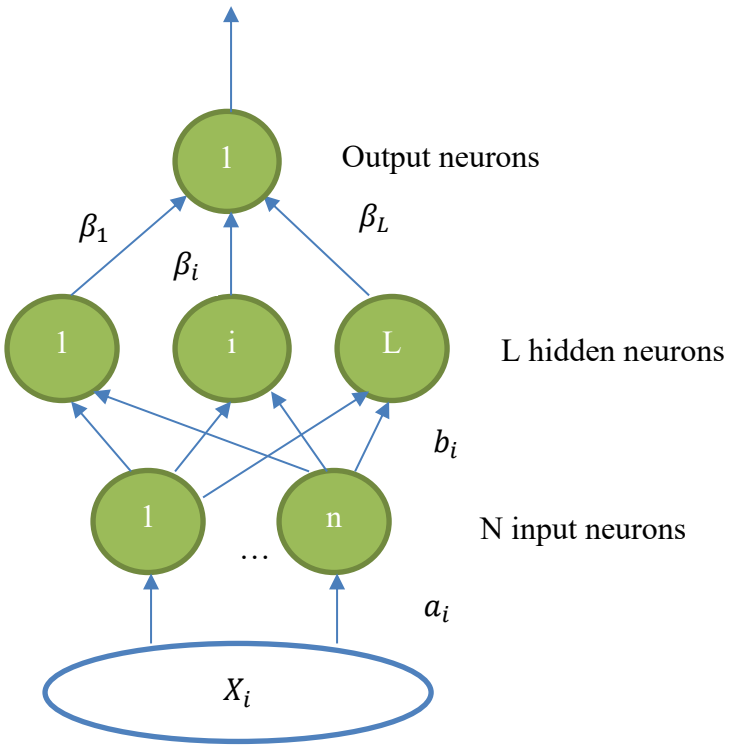


Figure 5: Single layer neural networks [34]

In general, ELM maps any given SLFN hidden layer into a matrix form:

$$h(x) = [G(a_1, b_1, x_1), G(a_2, b_2, x_2), \ \ldots. \ , G(a_L, b_L, x_L)] \qquad \ldots\ldots\text{Equation 8}$$

where a, b are the random initialized hidden layer parameter matrix, L is the number of nodes in hidden layers, G is the node activation function, which could be additive, radial basis function (RBF), etc. The additive node activation has the form:

$$G(a_i, b_i, x) = g(a_i x + b_i) \qquad \ldots\ldots\ldots\ldots\text{Equation 9}$$

where a_i is the weight vector connecting the ith hidden node and the input nodes and b_i is the bias of the ith hidden node.

The RBF node has the form:

$$G(a_i, b_i, x) = g\left(b_i \left|\left|x - a_i\right|\right|\right) \quad \text{.....Equation 10}$$

Where a_i is the center of the ith hidden node and b_i is the impact factor of the ith hidden node. Therefore, the output function of the SLFN can be written as

$$f(x) = \sum_{i=1}^{L} \beta_i G(a_i, b_i, x) = h(x)\beta \quad \text{.....Equation 11}$$

Where *h(x)* is the hidden layer output corresponding to input sample *x* and β is the output weight vector between the hidden layer and the output layer. With calculated hidden layer matrix of N input samples:

$$H = [h(x_1)^T, h(x_2)^T, \ldots, h(x_N)^T]^T \quad \text{.....Equation 12}$$

And the target matrix:

$$T = [t_1, t_2, \ldots., t_N]^T \quad \text{.....Equation 13}$$

Then β can be calculated as:

$$\beta = H^{\dagger}T \quad \text{.....Equation 14}$$

In this way, the input layer and hidden layer parameters a_i, b_i do not need to be tuned and the network can be trained very efficiently.

In particular, for a binary case, the decision function of ELM classifier can be written as,

$$f(x) = sign(h(x)H^T \left(\frac{I}{C} + HH^T\right)^{-1} T) \quad \text{..... Equation 15}$$

where H is the hidden layer matrix calculated from the training samples, T is the target matrix of training samples and $\frac{I}{C}$ is a positive constant matrix for a stabler inverse result.

Extreme learning machine (ELM) is an efficient single-hidden-layer feedforward neural network (SLFN) which has generally good performance and fast learning speed. It has been shown as an effective learning method in a wide variety of applications. The efficiency of ELM provides the possibility to incorporate part-based model into naturalistic driving data process which normally has high requirement for speed. And also ELM can be used as a unified learning platform which does not need to tune the hidden layer parameters as traditional neural networks do. Thus in our work, a detection method based on ELM classifier exploring HOG feature is proposed to achieve a better detection rate improving the processing speed.

3.6 Block Diagram Of Our Methodology

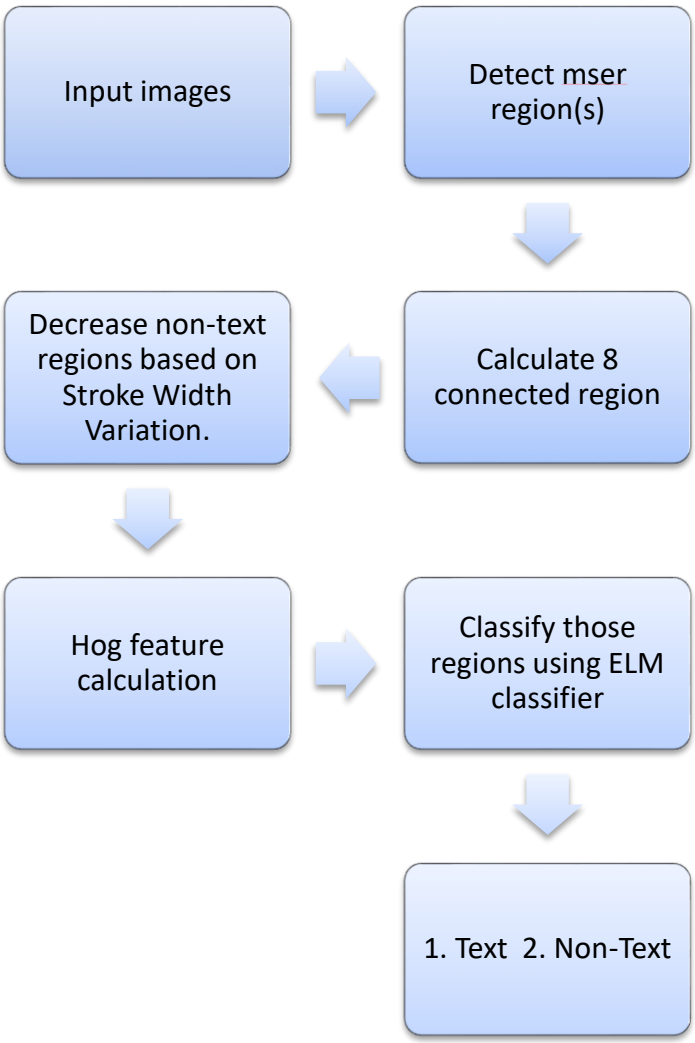


Figure 6: Block diagram of our proposed methodology

CHAPTER FOUR

EXPERIMENTAL RESULT AND ANALYSIS

In the present work, an Extreme Learning Machine (ELM) based classifier has been designed for localization of text contents in natural scene images using the Histogram of Oriented Gradients (HOG) feature on the benchmark dataset ICDAR-2015 which consists of *229 training images* and *233 testing images*. One of the freely available open source ELM tools called simple ELM classifier [43] has been used. Here, we have used 180 number of hidden neurons and sigmoidal function as an activation function for this classification technique. The detailed description is given in the chapter-3. The technique has obtained a recall of *95.39 percent* on the test data set.

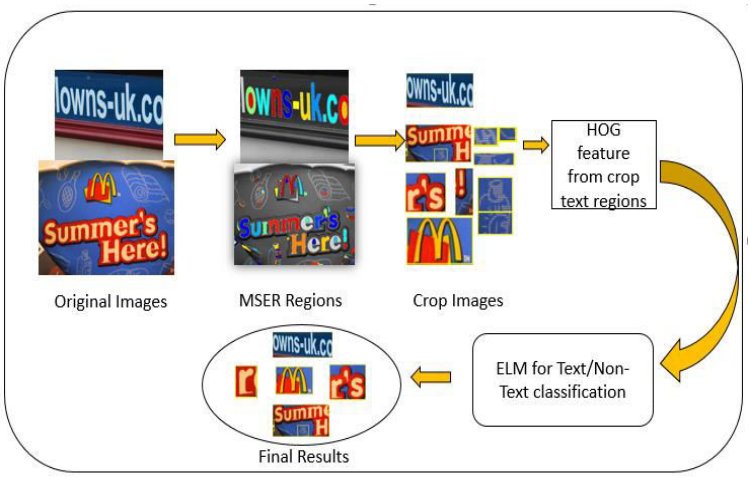


Figure 7: Block diagram showing intermediate steps in our proposed work

Figure 8: a) Original input RGB images, b) Localizes text region over the original images

Figure 9: a) Original input RGB images, b) Localizes false region over the original images

Figure 10: a) Individual text characters and (b) detected merged word or text lines

Table 4.1: Comparative Study of Results on ICDAR-2015 Dataset

Works	*Precision (%)*	*Recall (%)*	*F-measure (%)*
Feng et al. [35]	67.4	69.7	68.5
Yan et al. [36]	74.6	73.9	74.25
Tian et al. [37]	74.0	52.0	61.0
Hyun et al. [38]	51.88	47.52	49.60
Proposed method	94.02	95.39	94.70

Figure 7 is a block diagram illustrating intermediate steps of our methodology for text localization on the original images. Figure 8(a) and 8(b) shows example of some input images and localized text region on those input images. Figure 9(a), 9(b) describes example of original image and falsely detected region of interest where our method could not find text region. Figure 10(a), 10(b) illustrates an example of the localized individual characters on the input image and join the nearby region to localize a whole word region not a single character. Table 4.1 makes a comparative study of results of different methods on ICDAR-2015 dataset.

CHAPTER FIVE

CONCLUSION

Text localization and recognition from scene images is a complex Computer Vision task that is being studied by many research laboratories and international companies for its importance and critical use in newly developed technologies, such as automated driving and automated indexing of information from visual data. Unfortunately, till now no method proposed in literature achieves text localization and recognition rates that are even remotely comparable to human observers' performances. For this very reason, competition among different text localization and recognition methods in this field is still very strong, especially on standard datasets like the ones from the International Conference on Document Analysis and Recognition (ICDAR) or the ones for real-world applications from Google, such as Google Street View House Numbers (SVHN) for house numbers localization and recognition from images harvested from Google Street View; Google Street View Text (SVT) for cropped lexicon-driven word recognition and full image lexicon-driven word detection and recognition from Google Street View. As described in this thesis, we have presented here an approach for text localization from natural scene images using Extreme Learning Machine. This method achieves state-of-the-art performance on a benchmark dataset ICDAR-2015. In Chapter- 3 and Chapter- 4 we exploit Maximally Stable Extremal Region (MSER) to obtain state-of-the-art accuracy rates for text localization from scene images. In this solution, stable connected components are not discarded on the basis of their geometric properties; this assures that uncommon text fonts that are typically filtered out as noise elements by competing approaches are correctly retained and identified.

5.1 FUTURE WORK

In the last two decades, scene text detection methods have evolved tremendously, utilizing several vision techniques with increasingly powerful classifiers. The state of the art approaches are capable of detecting and recognizing English text with sufficient accuracy. In my opinion, every possible improvement to the presented methods of Chapter- 3 and Chapter- 4 requires the introduction of deep architectures and the subsequent collection of more labelled training data. In fact, deep-based approaches are currently the most effective and innovative way for reaching good results for text localization and recognition in natural scene images. The positive side of these novel deep architectures is that their use allow text localizing and recognizing methods to reach significantly higher detection rates than traditional shallow approaches. The negative side is that deep models require millions of training samples to reach human-level performances. In fact, as also stated by many experts in text localization and recognition fields, the best way to improve state-of-the-art results for deep-based text spotting algorithms consists in designing and developing a system for the automatic generation of synthetic train images that accurately simulate/mimic all the natural text elements and difficult conditions commonly found in real-world natural scene images.

CHAPTER SIX

REFERENCES

[1] C. Republic, “Real-Time Scene Text Localization and Recognition,” 2012, 25th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012, June 16-21, Providence, RI, USA.

[2] Neumann, Lukas, and Jiri Matas. "Text localization in real-world images using efficiently pruned exhaustive search." *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011.

[3] Bigorda, L., and Dimosthenis Karatzas. "A Fast Hierarchical Method for Multi-script and Arbitrary Oriented Scene Text Extraction." *CoRR* (2014).

[4] Gomez, Lluis, and Dimosthenis Karatzas. "Multi-script text extraction from natural scenes." *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013.

[5] Jaderberg, Max, Andrea Vedaldi, and Andrew Zisserman. "Deep features for text spotting." *European conference on computer vision*. Springer, Cham, 2014.

[6] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Reading Text in the Wild with Convolutional Neural Networks,” *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.

[7] Y. Zhu, C. Yao, and X. Bai, “Scene text detection and recognition : recent advances and,” vol. 10, no. 1, pp. 2015–2016, 2016.

[8] Yang, Kai, et al. "An extreme learning machine-based pedestrian detection method." *Intelligent Vehicles Symposium (IV), 2013 IEEE*. IEEE, 2013.

[9] Pan, Yi-Feng, Xinwen Hou, and Cheng-Lin Liu. "Text localization in natural scene images based on conditional random field." *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009.

[10] Wang, Kai, Boris Babenko, and Serge Belongie. "End-to-end scene text recognition." *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011.

[11] Matas, Jiri, et al. "Robust wide-baseline stereo from maximally stable extremal regions." *Image and vision computing* 22.10 (2004): 761-767.

[12] Tian, Shangxuan, et al. "Scene text segmentation with multi-level maximally stable extremal regions." *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014.

[13] Neumann, Lukas, and Jiri Matas. "A method for text localization and recognition in real-world images." *Asian Conference on Computer Vision*. Springer, Berlin, Heidelberg, 2010.

[14] C. N. Nafla, K. Sneha, and K. P. Divya, "Scene Text Detection Using Machine Learning Classifiers," vol. 4, no. 5, pp. 601–605, 2015.

[15] Epshtein, Boris, Eyal Ofek, and Yonatan Wexler. "Detecting text in natural scenes with stroke width transform." *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010.

[16] Feng, Yingke, Jinmin Zhang, and Siming Wang. "A new edge detection algorithm based on Canny idea." *AIP Conference Proceedings*. Vol. 1890. No. 1. AIP Publishing, 2017.

[17] Huang, Weilin, et al. "Text localization in natural images using stroke feature transform and text covariance descriptors." *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013.

[18] Risnumawan, Anhar, and Chee Seng Chan. "Text detection via edgeless stroke width transform." *Intelligent Signal Processing and Communication Systems (ISPACS), 2014 International Symposium on*. IEEE, 2014.

[19] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. Springer, Cham, 2014.

[20] Coates, Adam, et al. "Text detection and character recognition in scene images with unsupervised feature learning." *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011.

[21] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, “Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks,” pp. 1–12.

[22] C. V Dec and K. Simonyan, “Natural Scene Text Recognition,” pp. 1–10.

[23] Neumann, Lukáš. "Scene Text Localization and Recognition in Images and Videos." (2017).

[24] de Campos, Teo, Bodla Rakesh Babu, and Manik Varma. "Character recognition in natural images." (2009).

[25] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, “ICDAR 2003 Robust Reading Competitions,” no. Icdar, 2003.

[26] C. Yao *et al.*, “Incidental Scene Text Understanding : Recent Progresses on ICDAR 2015 Robust Reading Competition Challenge 4,” vol. 4, pp. 1–3, 2015.

[27] “Region detectors Requirements for region detection.”

[28] A. Vedaldi, “An Implementation of Maximally Stable Extremal Regions,” pp. 1–7, 2007.

[29] A. Gosavi, A. Gurav, and G. Bisht, “Text Detection and Translation,” vol. 6, no. 4, pp. 3306–3309, 2016.

[30] Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, 2005.

[31] Opitz, Michael. *Text Detection and Recognition in Natural Scene Images*. na, 2013.

[32] E. Ahmed, G. Shakhnarovich, and S. Maji, "Knowing a good HOG filter when you see it : Efficient selection of filters for detection," no. i, pp. 1–15.

[33] G. Huang, S. Member, H. Zhou, X. Ding, and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," vol. 42, no. 2, pp. 513–529, 2012.

[34] G. Huang, Q. Zhu, and C. Siew, "Extreme Learning Machine : A New Learning Scheme of Feedforward Neural Networks," *IEEE Int. Jt. Conf. Neural Networks*, vol. 2, pp. 985–990, 2004.

[35] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-Oriented Text Detection with Fully Convolutional Networks," *Cvpr*, pp. 4159–4167, 2016.

[36] J. Yan and X. Gao, "Detection and recognition of text superimposed in images base on layered method," *Neurocomputing*, vol. 134, pp. 3–14, 2014.

[37] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9912 LNCS, pp. 56–72, 2016.

[38] J. H. Seok and J. H. Kim, "Scene text recognition using a Hough forest implicit shape model and semi-Markov conditional random fields," *Pattern Recognit.*, vol. 48, no. 11, pp. 3584–3599, 2015.

[39] https://en.wikipedia.org/wiki/Google_Glass,2016.

[40] http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/. 24

[41] http://rrc.cvc.uab.es/. 25, 26

[42] http://vision.ucsd.edu/~kai/svt/. 27

[43] http://www.ntu.edu.sg/home/egbhuang/elm_codes.html

[44] https://vision.cornell.edu/se3/coco-text-2/

[45] http://www.iapr-tc11.org/mediawiki/index.php/KAIST_Scene_Text_Database

[46] http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/IIIT5K.html

[47] http://www.iapr-tc11.org/mediawiki/index.php/NEOCR:_Natural_Environment_OCR_Dataset