A Hybrid Approach to Detect Text and to Reduce the False Positive Results in a Scenery Image

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Technology

in

Computer Science and Application

Submitted By

Animesh Sharma

(Roll No. 601403005)

Under the supervision of:

Dr. Rajiv Kumar

Assistant Professor

Thapar University, Patiala

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

June 2016

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*A Hybrid Approach to Detect Text and to Reduce the False Positive Results in a Scenery Image*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Application* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Rajiv Kumar* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Animesh Sharma

(Animesh Sharma)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Rajiv Kumar

(Dr. Rajiv Kumar)
Assistant Professor,
CSED

Countersigned by

(Dr. Maninder Singh)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. S. Bhatia)
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life. It is a great privilege to express my gratitude and admiration towards my respected supervisor **Dr. Rajiv Kumar**, Assistant Professor, Computer Science & Engineering Department, Thapar University, Patiala. He has been an esteemed guide and great support behind achieving this task. This work would not have been possible without the encouragement and able guidance of him. I also thank my supervisor for his time, patience, discussions and valuable comments. His enthusiasm and optimism made this experience both rewarding and enjoyable. I am truly grateful to him for extending his total co-operation and understanding whenever I needed help and guidance from him. I am also heartily thankful to **Dr. Maninder Singh**, Associate Professor and Head, Computer Science & Engineering Department and **Dr. Sanmeet Kaur**, PG coordinator, for motivation and providing uncanny guidance and support throughout the preparation of the thesis report.

I will be failing in my duty if I do not express my gratitude to **Dr. S. S. Bhatia**, Senior Professor and Dean of Academic Affairs, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar University memorable. Last but not least, I would like to thank my family for their wonderful love and encouragement, without their blessings none of this would have been possible.

Animesh Sharma

(601403005)

Abstract

Detection of the text from the scenery images containing text is a challenging task that has received a lot of attention recently. In scenery images there are two key components 1) finding text from images, 2) Recognition of character. Many researchers have published their work on both components. Finding the text in the images is the primary part because the overall accuracy of the model depends on the output of this phase. In the present study, a method has been proposed that consists of two phases 1) Text detection 2) Text verifier. Text detection is done through the well-known algorithm for text detection-MSER (Maximally Stable Extremal Regions) feature detector. Then different filters like elimination of non-text region based on simple geometric properties, and elimination of non-text region based on stroke width variationon the output of MSER feature detector are applied to filter out the components that possibly cannot be the text. In second phase, machine learning approach (ANN-classifier which acts as text verifier) is used to classify the text and non-text on the final output of phase 1. It is found that proposed algorithm almost eliminate all the false positive results on the final output of the MSER feature algorithm.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Text plays very important role in the scenery images and video frames, understanding that text carries important information is the challenging task. Researchers were found that human pay more attention to the text rather than the object in the images. Actually text detection is the area in which one tries to find all the text area in the scenery images and video frames by finding precise borders of text lines through various approaches. Text detection and localization phase is the important phase because the subsequent phases (segmentation and text recognition) all depends on the output of the text detection and recognition phase. In this thesis the word *scene text* is used to refer the scenery image and video frames containing text. This chapter covers the introduction of the problem and challenges in scene text, then difficulties and characteristics in scene text, then relationship between scene text and other field, and at the end brief history of scene text.

1.1. Brief Introduction of Scene Text

Day by day the number of users of portable camera, video recorders and camera phone are increasing and this leads to large growth in multimedia data in the form of images, videos since from 1990 [1-4]. As a result, there is a requirement for the efficient indexing and retrieval process for the huge multimedia databases. Because of this need researchers are working on this area with a big question how to do this efficiently. Additionally there is a challenge to tag and interpret events logged in the video the basis of semantics. The content-based image indexing is the method of assigning tags to video or images founded on their insides. To do this one need to find the semantics of the image in the form of face, text, human action, and vehicle. Among them, text in the scenery images and video frames are the primary concern. Texts are very helpful in analyzing the insides of videos or the images, also it is stress-free to extract compare to extra features of semantic subjects, and it is favorable for applications like spontaneous video logging and text-based image search. It is also beneficial for the car navigation system. The car navigation system can send the alert message to the driver according to the street sign board. This method can also support the visually impaired people in the streets, for getting the direction.

The traditional OCR (optical character recognition) system that is used for the document recognition, in the early 1960soptical character recognition was some of the major applications of pattern recognition. Now a day, for the spotless and well-organized text, document recognition is seen as nearly a cracked problem [5]. A number of works is done on the traditional OCR system and because of this research; it is easy to recognize the text from the scanned document that has low noise. In the current world, the process of transformation of more and more data into digital form is going on. Visual texts appear in many forms like in scenery images or in videos. The problem with traditional OCR system is that it does not work well in case scene text. When the scene text is fed into the traditional OCR system the performance of the system degrades typically ranges from 0% to 45%, it is because the nature of scene text is very different from the traditional document analysis.

Associated with the text in the documents, the text in the scenery images/video frames is in far smaller amount. The text in the video frames gives the critical information about the type of the media content. They appears in the form of products, locations, brands, names, score of match, date, and time which gives the interesting info of the type of video insides. The scene text can be of different colors, fonts, sizes, alignment, movement; lightning condition and can superimposed on the different backgrounds or fixed on the exterior of different objects. Therefore finding and extracting the text is difficult job in scene text. The main focus of the researcher is to find the best way of finding and extracting the different types of text in scenery images or in video frames having complex backgrounds. One can say that it is extending the application of OCR system into broader area and also service people for additional understanding the machine of the text detection and recognition. There are different types of method, such as texture-based, region-based method, and some other techniques [1-13], have been suggested. Many of these techniques have attained the remarkable performance.

Figure 1: Example of Scenery Image Containing Text

Figure 2: Example of Video Frames Containing Captions

The text in the video can be classified into two types. The first one is scene text that is naturally occurring in the video frames or in scenery image as given in the Figure 1. The second one is caption text in which the editor falsely superimposed the text on the video as given in Figure 2. From the examples, it is clear that scene images containing text is more puzzling than caption text, and it is because scene texts have the varying structure, lightning effect, transformation, and complex movements. Text detection and extraction from the scenery image or in video frames are exciting problem for the research community and it is because of the subsequent deviations in the properties of the scenery image or video frames.

- Size: The size of the text can vary a lot, different kind of assumption should be based on the different application.
- Alignment: The text in the video frames like caption text mostly present in the horizontal orientation but sometimes it can appear in non-planar way because of the special effects. In scenery image the text can be ranged in several orientation and also can have different structures and geometrical spins.
- Color: Connected component based method can be used for the text detection if the colors of all the texts are same. The most research done until now is mainly focused on finding the text strings of the same color which is also known as monochrome but the video frames and other scenery images may contain the texts in different colors, i.e. More than two color which is also known as polychrome.
- Edge: All the scene text is aimed to be in readable format. Because of this there is strong edge between the boundaries and background of the text.
- Compression: In the current world, different images videos are recorded, shared, and handled in a compressed setup. Dealing with the compressed setups is not an easy task.

From the above points it is clear that text extraction from video frames as well as scenery image is not a simple task. Researchers proposed many methods to solve this problem that generally has the following steps:

- Text detection, searching for the text in video frames or scenery image.

- Text localization, grouping of all neighbor text into one region and creating the tight bounding boxes all around the text regions.
- Text binarizations, the text lies in the bounding box are binarized with same pixel value. (1 for text and 0 for background or vice versa).
- Text recognition, at the end performing the OCR on the binarized text image.

Figure 3 shows a block diagram of different stages mentioned above, in which the primary focus of research authority is in the text detection and text localization phase, because the overall accuracy of the whole system will much depend on these two stages.

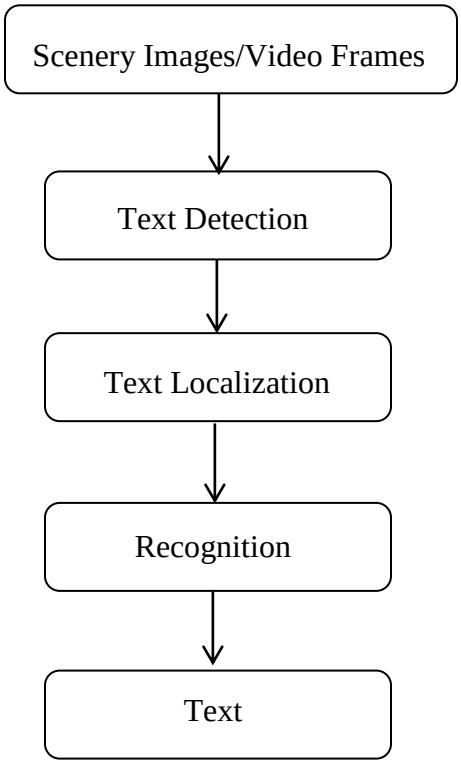


Figure 3: Block Diagram of the Detection and Recognition of Scene Text.

1.2. Difficulties and Characteristics in Scene Text

A video frame may contain the scene text or caption. Since captions have good contrast and clarity and is mostly in the horizontal orientation with less noise. Therefore to find the caption in the video frames, one can take the advantage of the same properties of traditional OCR system to some extent:

- Text will always appear in the forefront.
- Mainly this text is projected with suitable light condition that makes the text independent from the scene.
- Text pixels values are distributed accordingly to limited rules.
- Most of the time texts are aligned horizontally.
- Font, size, orientation, and spacing are constant inside the text area.
- Text pixels have the greater contrast related to the background.
- There is an identical color for all the text in same text area.
- In most of the cases the text will appear in the bottom the video frames.
- Texts are clubbed as the group of words rather than isolated characters.
- The background is typically identical for the moving text.
- The text is typically moving either in vertical or in the horizontal direction.
- Similar text may appear in different consecutive frames.
- Generally texts have the low resolution.

While the scenery image containing text reveals the following properties.

- Text can be appearing within objects or with the noisy background.
- There is lot of difference in the size, color, font, style, orientation, and alignment even inside the similar word.
- There is a chance that portion of the text excluded because of the camera and object movements.
- May there is complex movement in the scene.
- There is lot of lighting variation.
- There is a chance of deformation if text appears in the stretchy surface like cloth.

- Text may appear with the special arts or effects and with some images.

From the above mentioned characteristics of the video frame containing text, it is clear that one have to apply different strategies for different inputs. Some Method will work fine on video frame containing the captions but not on the video frames containing the scenery text.

Different types of method have been discovered on the basis of contrast and texture of text. By describing the text as some different texture, the problem of extracting the text from complex background and low contrast can be reduced, but as long as the image contains the high value of contrast and have smooth pattern built on the forms of the character components, these techniques produces the decent outcomes. For the scene text containing complex structure, different fonts, and structure in background that is as like as textis still an interesting problem.

There is also a machine learning approach to answer this problem, in which one need to train the classifier by using labeled data or unlabeled data, to differentiate the areas of scene text as text and non-text area, but the problem with theseapproaches is that it will use the big number of training data containing different kind of text and non-text area, so that it can handle all kind of variations, and this leads to additional computational burden. Also choosing the text and non-text training samples are not simple task.

Toenhance the accuracy of the text detection method one need to perform the preprocessing like increasing the contrast, and increasing the sharpness. Normal operation like minimum, maximum, median and averaging are not enough to improve the sharpness and contrast of the pixels which is a text, because above processes can cause blurness at the edges. Some techniques are discovered on the basis of edge strength and gradient evidence. These method works well because of the fact that edge deepness is strong at or nearby the edge of a character stroke. Because of the above assumption, edge based method is more powerful than the connected component based or texture based techniques. The problem with this algorithm is. Even though these algorithms are fast but the problem with this algorithm is that one has to define the gradient and edge strength for all the pixels in the unstructured background and also it can produce a lot of false

positives results as the output, it is so because the defined steps of algorithms may also be fulfilled by background components like windows of building tree leaves, branches, etc.

There are different researchers who suggested the techniques by joining some features of text such as edge strength and color using various classifiers, so that they can categorize between text and non-text. There is chances that this algorithm may give the best result, but the problem with this algorithm lie in searching which features gives the best result to categorize between text and non-text. Actually there is no any standard method for combining different classifiers.

Large amount of methods are proposed that usually assumes only horizontal text lines. They are ignoring the other orientation of the text because of the complexity to deal with different orientation in there algorithm. The main focus of the researchers and the algorithms is on clubbing the CCs by nearest neighbor and textual features standards. It works well only if there is space between the text lines. Otherwise these algorithms may finish up with clubbing two or more than two close text lines and the background into one area.

In summary, it can be said that there are many methods for text detection problem, but none of the discussed techniques is the perfect solution for the text detection system that has the noisy and complex unstructured background.

1.3. Relationship between Scene Text and Other Fields

Text detection and localization system contains different areas like, image processing, pattern recognition, artificial intelligence, multimedia, computer vision, document image analysis, and software definition to create software working and complete. One can say that the text detection in scene text is the extension of traditional OCR system in a way that after extracting the text region from image, the segmentation and the recognition phase is same as the traditional OCR system. Actually text detection phase is also same as the traditional OCR system because in traditional OCR system, the separation of foreground (text) from background is done, which is simple task since the text in document image contains the high contrast and high resolution with spotless background

but it is difficult for the text detection in scene text because scene text can have any contrast and resolution with complex background structure.

Whenever one considers the scene text then image processing is required, so that features can be extracted to identify text. All the discussed features and properties are mined via image processing tools like gradient computation, color analysis, Fourier coefficients analysis, edge detection, wavelet coefficients analysis, morphological operation, etc.

Now large volume of the researchers uses the recent methods in computer vision to extract the features scene text like SIFT, HOG, SURF, super pixels and MSER. According to the surveys, these features are steadier compared to the traditional features as they are vigorous compared to geometrical transformation and distortion. These features are currently becomes the trend in this area.

Feature extraction is significant when it does several tasks like classification and recognition. By using the pattern recognition based system one can perform the classification and recognition. Basically one can perform the above operation by using the supervised classifier like neural networks, Bayesian, SVM etc. and unsupervised classifier like discrimination analysis, clustering, etc. The feature in both the classifier extracted in the same way but in case of supervised learning, training data is required and in case of unsupervised learning, there is no any training data.

Real time systems are required for video text detection system. For example, by detecting and recognizing text from street sign board automatically, one can help the visually impaired people to walk on the roads accordingly. That means a complete system needs a graphical user interface and hardware support for the real time use.

1.4. A Brief History of Scene Text Detection

The origin of the research on video frame or scenery image containing text may be tracked from the document image analysis because of its core methods in optical character recognition (OCR). The use of document image analysis is actually about 40 years old. The designed printed digit which is an OCR font was first used in the business around 1950s. The first social security administrator device and postal address reader to

read the typewritten were installed in the 1965. In the 1980s, there is a machine to read simple hand printed forms and typeset materials.

Some of the initial phases of processing scanned documents were free from the type of the document. Binarization, noise filtering, segmentation and edge extraction methods can be applied on handwritten as well as the printed text. Segmentation have to be done in proper way, at first line wise then word wise then character wise once it is segmented in the proper components then one need the more specific techniques. Usually this field is mainly divided into areas. First one is handling of text documents and second one is handling of graphics document. The pages of the text documents are segmented into the columns, paragraphs, lines, words, and at the end characters. Then OCR translates each character into the text that is readable by machine (generally character code like Unicode, ASCII, andEBCDIC), but it is known that the documents not only contains some strings but also the tables, words in different fonts and effect. So to analyze the whole document all the above discussed parameters would be considered. Engineering drawing, music scores, maps, schematics diagrams, and chart of organization are the main examples of the graphic documents.

One can divide the approaches of document image analysis into two ways, a) top-down and b) bottom-up analysis. Most of the document image analysis follows the top-down approach in which one can first try to find the larger components like paragraphs and columns then the other small component like lines, words and characters. While in bottom-up analysis one can form the word first then word into lines and then lines into paragraph and so on.

Models have been established for equations, formulas, tables, business forms, circuit schematics, tables, mechanical drawing, flow charts, chess notations, and music symbols. Most of the above examples have the properties of natural language and remaining have the properties of the domain-specific constraints.

From the 1960s, lots of the research has been done based on the OCR. Number of the OCR systems was built for precise domains started appear in the marketplace. In the time period of 1980-1995, many techniques were suggested, and also many document analysis

systems were proposed. More than 500 papers were reported at the international conferences in only document analysis (ICDAR 91, ICDAR 93, and ICDAR 95).

New decades come with new challenges, since with the inventions of cheap storing devices, which can easily stores thousands of images in the form of compressed bitmaps. Researchers started work in translating them in machine searchable format. In 1999, the document imaging, i.e., which is an electronic document storage free of sophisticated image manipulation, is on the peek. Document image interpretation is considered as the small part of the document analysis. Therefore researchers found different applications in which they can extend this concept. Because of this extension in different applications, document image analysis becomes the influential way for transferring the knowledge. Actually most of the knowledge is acquired from the documents like government files, technical reports, books, newspaper, magazines, journals, bank checks, letters, etc. If someone will do this work manually then it will very time consuming, therefore automatic knowledge acquirement from any kind of document becomes the significant interest. One of the best applications where the document image analysis plays an important role is postal automation. Without OCR system it is very time consuming for the post office to read and arrange the packets accordingly. Know a days the system exists in postal office that will sort the letters accordingly. The system are well operated and established but the accuracy till reported is only 55% [5]. One of the main reasons behind this poor accuracy is the problem in finding the address block in the letters. If the parcels or letters are packed with some color paper or with graphical papers, then it is more challenging to recognize the address blocks. In most of the letters, the envelops are plain and the printed text is of different color in this case it is easy to recognize the address block. But if the printed text and the background are of similar color then it is more challenging.

There are several approaches suggested for solving the automatic localization of address blocks in mail letters. A well-structured review can be seen in [14]. In which a texture segmentation technique built on multichannel Gabor filters has been used effectively [15]. The similar technique is applied with some modification in [5] so that it can deal

with low resolution image, handwritten, and machine written address blocks. Finding the address block in the postal letter led to the discovery of new research in document image analysis, which is find the text in the complex background images like book cover, CD cover etc. The images that have the complex background then it is difficult to segment the characters by suing modest thresholding, where the size, color, font and the text orientation are unfamiliar. Consider the full page document which contains the text as well as the figures. Human mind can easily differentiate this but it is difficult to make a machine that can do this task, and because of this various methods are developed to do this task.

The best natural characteristics of text are there regularity. Usually printed text contains the characters that have the same color, size, and uniform stroke width. Characters usually have the same space from their neighbor characters. There structures are identical means they are of the same height in the lines, and the lines form the columns. So there are many methods were build which takes the advantage of these features and have the smooth performance. To extract the characters from the unstructured background, the connected component techniques have been used. At first text region should be detected, by using some techniques like connected component analysis and adaptive thresholding. Then one can apply more filters to discriminate between the characters and non-characters regions. Or one can directly apply the OCR phase but OCR that has the feature to ignore the non-text regions. Actually we can`t extract the text region without recognizing the text from the complex background. The performance of the OCR system will reduce if one directly feed the image (without segmenting the text region) with complex background in the OCR, and it is not reasonable at all because the image contains a lot of complex structure. So best way to deal with this is to first define the text and non-text region then ignore the non-text region and feed only the text region in the OCR to get the better performance. Because of this intention, Zhong et al. [16] proposed two ways of extracting the text. The first technique is based on the segmentation of the color images into connected components of uniform color and additional filters to reduce the false positive result. The second technique finds the text line rather than distinct characters by using the spatial gray value variance intimate in text area and the

background. Hybrid method is then proposed by combining these two approaches for image containing complex background structure.

In the last 40 years, the applications of the OCR have increased tremendously, and it now beyond the early work in the field of OCR. Now a days, the field has covered lot of different areas like physical and logical layout analysis, preprocessing, optical and intelligent character recognition, form processing, graphics analysis, writer verification, signature verification. The tremendous study in this area results in many practical applications, like forensics, office automation, and digital libraries. 20^{th} century is a digital century. Most of the people now a day carrying the camera phone, digital camera and have PC cams. A lot of information is collected from these devices. There is a survey paper on camera-based scene image [5] of 2003 in which they projected future interest of OCR community in video frame text detection system. Even though lot of research has been done in scenery images containing text but the performance of the existing algorithm is not up to the mark. So still it is the area of interest for the researchers.

1.5. Thesis Organization

Organization of thesis is structured into various chapters. Outline of each chapter is structured as below.

Chapter 1 (INTRODUCTION) gives the basic review of thesis topic i.e. starting with brief introduction, then difficulties and characteristics, then relationship between scene text and other field, and at the end brief history of scene text.

Chapter 2 (LITERTATURE SURVEY) summarizes the work done by many researchers to overcome the problem of detecting and localizing of the text in the scenery images and video frames.

Chapter 3 (PROBLEM STATEMENT) includes basic definitions describing the problem that has been chosen in this research, and objectives that are needed to be attained for successful completion of research work.

Chapter 4 (PROPOSED METHODOLOGY) provides new methodology that has been proposed with motive to reduce the false positive regions in scene text. All the core terms

are defined first, then a brief discussion about the various datasets that has been used. At the end of this chapter a brief discussion of the implementation has been given.

Chapter 5 (EXPERIMENTAL RESULTS) includes study that is performed and Comparative analysis shows that proposed method provides better results in term of precision, recall, and f-score. Results have been described in tabular form.

Chapter 6 (CONCLUSION AND FUTURE SCOPE) includes the brief description of what the problem is and measures that are achieved against objectives and scope for further research has been stated clearly under the future scope section.

Chapter 2 Literature Survey

Research that have been made in the field of text extraction is analyzed and methods that been used from past have been described in details. The method for finding the text from the scene text can be generally classified into two types: a) region based method b) texture based method. Finding the text in compressed domain is discussed in section 2.3. Experimental results and the performance measure of each approach are discussed whenever it is available.

2.1 Region-based Methods

Region based methods use the different kind of properties like gray-scale, color of a text area, and the difference between the color of the text region with the background. Region based methods are again classified into two types: a) connected component based method b) edge based method. Both methods are based on bottom-up approach; First it will identify the sub-structures like connected component or edge, and then it will combine the sub-structures and makes the bounding boxes for the text. There are also some methods that use the mixture of both connected component-based and edge-based methods.

2.1.1. Connected Component-based Methods

Connected component-based methods are based on the bottom-up approach in which clustering of small regions into larger regions should be done until all the regions are covered in the image. To do this some geometrical analysis should be done to combine the text regions using the spatial arrangement of the regions so that text regions can be identified and non-text regions can be filtered out.

Ohya et al. [17] proposed a method contains the four stages: a) binarization of the image is done through local thresholding, b) by using the gray scale difference most likely characters are identified, c) each characters are verified by matching the pattern of each characters with the stored dataset, and d) to update the similarities relaxation operation are performed. The proposed method were able to extract and recognize the characters, from the scenery images including then multi segment characters, under different sizes,

fonts, lightning condition, and positions. The problem with this method is that it uses the binary segmentation which is inappropriate for the video documents, because the video may contain various types of objects having different gray levels and may also have variations in lightning effect and high level of noise effect too. Furthermore this approach fails when there is a different text alignment and monochrome texts. Based on the dataset containing 100 images, the recall rate for text detection was 85.4% and the character recognition rate was 66.5%.

Lee and Kankanhalli[18] proposed a method based on connected component for the detection and recognition of texts on cargo containers, that can have a different lightning conditions and the different sizes and shapes of the character. For the connected component generation edge information is used. To find the components which is most likely to be text, boundaries of the text determined by using the difference between the adjacent pixels value. Based on the pixels value in the boundaries, local threshold values are selected for each text candidate. Most likely characters are used to make the connected component with the similar gray level. Furthermore some different techniques are used to filter out the non-text regions based on the contrast histogram, aspect ratio, and run length measurement.

Zhong et al.[16] proposed a method that uses the color reduction technique. By using the peaks in the color histogram in the RGB color space they quantize the color space. It is based on the assumption that the most likely text areas will group together in the color space and also it will occupy a significant area in the image. Then each character component is passed through several different filters so that non-text regions can be eliminated and the filters are based on the diameter, area, and spatial alignment. The proposed method is then tested on book covers, and CD covers image.

Image segmentation was done in a color histogram by making use of color clustering in the RGB space by Kim[1]. In this procedure elimination of the image boundaries and removal of lines that are horizontal was performed. Then iterative projection profile was used for analyzing and based on it horizontal lines and text segments were extracted. After that these text segments are merged by making use of heuristics. However this

approach is not suitable because the threshold values has to be found out emphirically. An approximate of 60 videos images of different character size and styles were used for the experimentation purpose and around 87% localization rate was found out.

Text areas have homogeneity intensity in images. This concept was used by Shim et al.[19]. Merging of pixel with same grey level within a group is done. Then removal of large regions for the purpose of text regions sharpening was done by making use of boundary analysis. Verification of the candidate regions according to the contrast, area, fill factor and size is done. For the extraction of the text strings neighboring text areas are considered.

Text areas are regarded as CC and enhancement of the text extraction by applying motion analysis. Segmentation of image that is given as input according to the monochromatic property of text regions based on split and merges technique. Filtration of regions based on too small and too large has been carried out. The enhancement in the obtained results was performed by motion information, contrast analysis etc. the estimation of the motion was made by making use of absolute difference criteria of the block matching technique. The blocks that are missed out are discarded. The main focus is in those areas, which are likely to be text as they exhibit high contrast. The contrast difference between foreground (text) and background helps in filtering out the non-text components. Some more heuristics techniques are used to filter out more non-text components based on width, height, geometric properties and aspect ratio. Based upon the dataset containing 2247 frames, the proposed method extracts 86% to 100% of caption text. Figure 4 shows the different steps of the proposed algorithm when applied on the video frame.

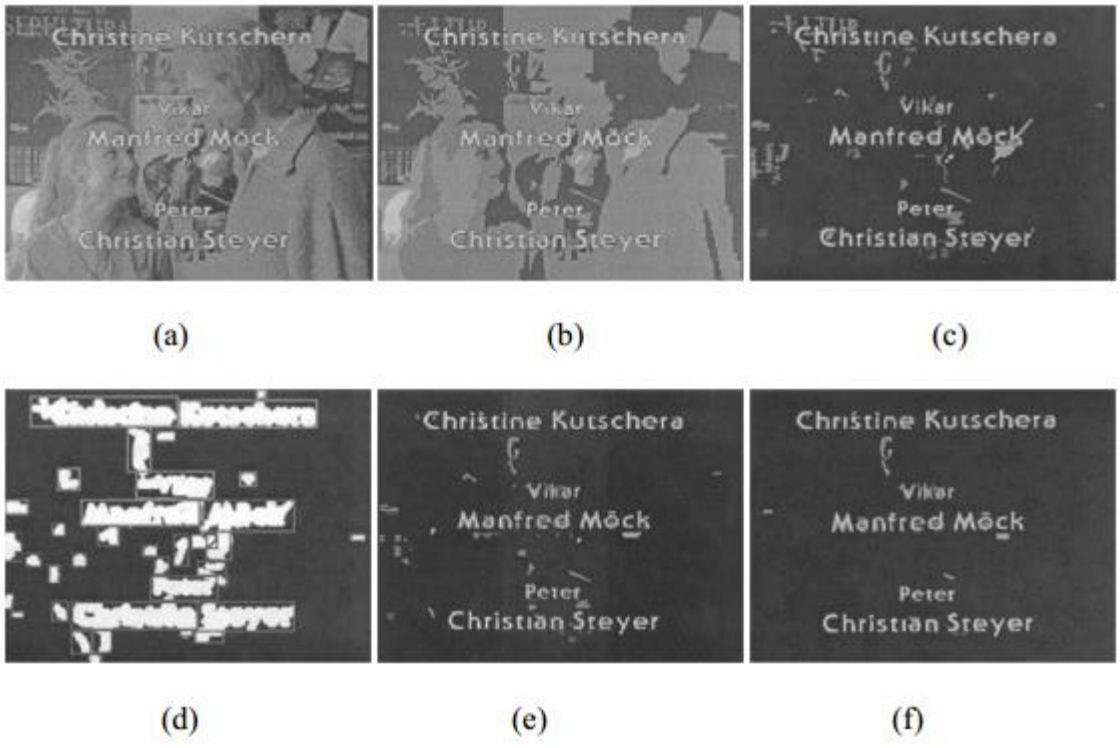


Figure 4:In-betweenSteps ofProcessing in the Techniqueby Lienhartet al. (a) OriginalVideo Frame, (b) Image Segmentation Using Divide and Conquer Method, (c) After Size Constraint; (d) After Binarizationand Dilation; (e) After MotionStudy; and (f) After Contrast Study and Aspect RatioConstraint.

Jain and Yu [20] proposed a technique based on the connected component. They applied the connected component technique after the preprocessing stage that contains several steps a) bit dropping, b) color grouping, c) multi-valued image decomposition, d) forefront image generation. A 6-bit image is generated by dropping the bits from the 24-bit color image, and then color grouping algorithm is used to quantize them. Decomposition of images into many forefront images was carried out and then each forefront image is passed through the same text detection stage. The decomposition of the multi-valued image is shown in Figure 5.Thenconnected component are generated for all the forefront images by using the technique called as block adjacency graph. Then all the detected text regions are in the different forefront images are then grouped into the output image. The proposed technique was tested in the different kind of images containing like

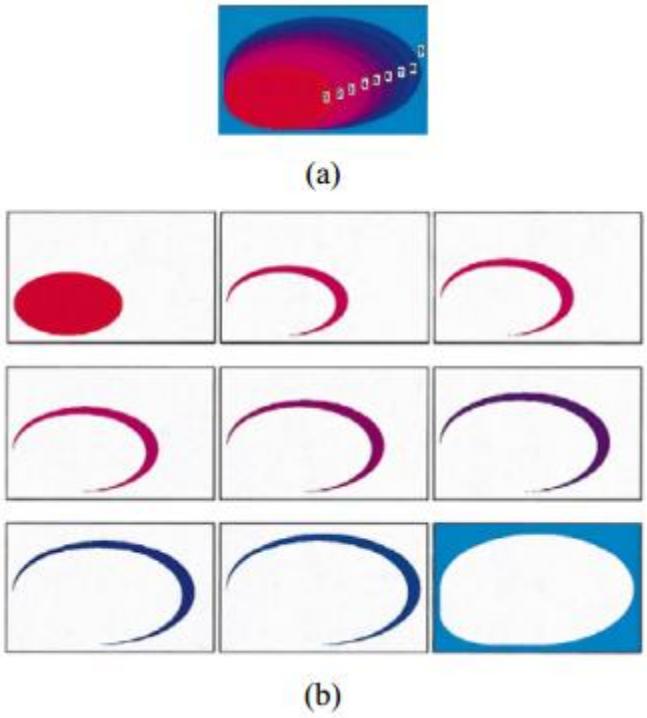


Figure 5: A Multi-colored Image and Its Element Images: (a) Color Input Image; (b) Nine Element Images

video frames, scanned color images, web images, and binary document images. The processing time was less than 0.4 second when it is executed in UltraSPARC I system having 64 Mb memory for the images of resolution 769×537. The problem with this approach is that it cannot be applied on the text which is skewed in nature. It works well with the horizontal and vertical texts.

Messelodi and Modena`s [21] proposed the method containing three stages: a) All the elementary objects are extracted, b) objects are then filtered out, and c) then text lines are selected. Before applying the actual technique all the images are first preprocessed. Then they applied the intensity normalization, image binarization, and at the end connected components are generated. After that different kind of methods like aspect ratios, sizes, areas and contrast are used to eliminate the non-text regions from the images. According to the author threshold value will depend on the different applications. At the end extraction of text a line were done starting from the single region and then expands recursively until some termination condition are not satisfied. The proposed algorithm

also works well with the partially skewed images. The technique was tested on 100 book cover images, and the accuracy for the text detection of 91.2 was achieved.

Kim et al. [22] filters out the non-text components by suing the cluster-based templates. Same kind of the approach is also used by Ohya et al. [34]. They also use the geometrical properties like area, size and alignment along with cluster based template to filter out non-text component.

Hase et al.[23] uses the connected component method for the color documents. They are having the same assumption for the text area detection algorithm that all the character is in the single color. Color histogram was created on the basis of the representative color (L*a*b color space are created by using the pixel value). Then several binary image is created from the images and by using the multi-stage relaxation string extraction was done. Then by using the conflict resolution rules, all text are fetched by their likelihood and this process was done after merging the all the outcomes from the different binary images. To find the likelihood of each character string, the standard deviation of the lines, width of the initial elements, alignment of the character string, and mean area ratio of the black pixel in elements to its bounding box are used. Filtration of the character string was done by using the two conflicts (overlap and inclusion) through the tree representation. The proposed can deal with shadowed and curve strings, but the problem with this algorithm is that the likelihood of a character is not easy to define and because of this character string can be missed.

By analyzing all the proposed method based on the connected component method one can say that this method actually have the four stages: a) Preprocessing (such as noise reduction, color clustering), b) connected component generation, c) Filter out the components that are not likely to be text, d) grouping of the different text components. The accuracy of the connected component based method is somehow affected by the grouping of the different text components, such as text lines selection or projectionprofile analysis. Furthermore to filter out the non-text component several threshold values are used, and generally threshold values are dependent on the dataset.

Weinman et al. [24] used a cluster of filters to study the texture topographies in different block and joint texture distributions between neighbouring blocks by using the conditional random fields. The limitation of these methods is that they used non-content-based image panel to split the image into blocks of same size before clustering is achieved. Non-content-based image divider is very prospective to breakdown the text characters into fragments that fails to gratify the texture restraints

Epshtein et al. [25] designed a content-based partition called as stroke width transform to excerpt texts characters with stable stroke widths. To catch the value of the stroke, width is calculated for all image pixels, and validates its use on the job of scene text detection in scenery images. This work is quite interesting because it can concurrently detect texts with different scale and is not limited to the horizontal texts only. As per the author the processing time taken by this algorithm is 0.94 second, and the precision of 0.73 and recall of 0.60 on the ICDAR dataset.

X.Huang et al. [26] in 2015 used two methods namely Maximally Stable Extremal Regions (MSER) and Support Vector Machine (SVM) for the development of a new text detection method. This method is robust and performs good content text analysis of the images. The recognition process in this method is divided into two phases. One phase belongs to the detection and other phase belongs to the recognition procedure. MSER and color clustering has been used in the detection phase for the extraction of the text from the scene images. After that the non- text regions are filtered. In the last stage with the help of text generation the word image is found. In the recognition phase segmentation of the word image takes place and character is recognized using SVM. The results are obtained on applying this method on a standard dataset. It has been observed that this method performs better than the conventional text detection and recognition methods.

2.1.2. Edge Based Method

Edge-based method uses the properties that there is a high contrast between the text and the background. After identifying the edges of the text boundary they are merged. Then by using some more filters the non-text components are eliminated. Generally the edge

detection is determined by the edge filter- Canny operator. After that morphological operator and smoothing operation are used for the merging state.

Smith and Kanade [27] they use the 3×3 horizontal differential filter to find the vertical edges based on thresholding on an input image. Then elimination of the small edges was done through smoothing operation. Bounding box is defined after the connection of adjacent edges. Then different heuristic techniques are used to eliminate the non-text regions like aspect ratio, size of the bounding box, and fill factor. In the end to filter out the groups that have the same texture and shape property, the intensity histogram of each group is analyzed. There is a difference between the method given by Sato et al. [28] and some other edge-based methods is in their usage of recognition based on the character segmentation. Position and segmentation of individual characters are directly depended on the character recognition outcome. It is because the decisions are made on the basis of the character recognition outcome.Therefore the performance for the character segmentation was improved. By merging the above process with the Smith and Kamade`s text detection approach , it was found that the whole system is taking less than 0.8 seconds for the dataset containing images of resolution 352× 242.

Hasan and Karam [29] proposed the method that is based on the morphological approach to detect the text from the image. To achieve this RGB components of color image are combined and the resultant value Y gives the intensity of the image and the equation 1 to find the Y is given below:

$$Y = 0.299R + 0.587G + 0.114B \quad (1)$$

The proposed technique is now used by the many researchers to deal with color images because of its simplicity. But the problem with this approach is that when color image is converted to gray scale image then there are chances that different components may have same grayscale value even though they have different color. To illustrate this color image and grayscale image are given in Figure 6.

(a) **(b)**

Figure 6: Difference between Color Image and Its Gray-scale Image: (a) Color Image, (b) Corresponding Grayscale Image.

After this, edges are identified by using the morphological gradient operator. Then by using the threshold value binary image is generated.In each candidate area in the intensity image adaptive thresholding is performed. Candidate regions are then formed by using the edges that are spatially closed. And the small components are eliminated by erosion. Then some more non-text components are eliminated by using size, aspect ratio, thickness, and gray-level homogeneity. From the experiment they done it is found that the proposed technique is robust to noise. But the problem with this approach is that it will fail when the text in the image are skewed.

Chen et al. [30] proposed the method based on the edge detection through canny operator. To reduce the computational complexity, estimation of orientation and scale is done through only edge point in a minor window. To connect the edges into the clusters morphological dilation is performed. Then some more heuristic technique like aspect ratio and height is applied to filter out non-text clusters. To generate input features for scale estimation two groups (stripe form filer and edge form filter) are used based on Gabor-type asymmetric filters. To do the approximation of the scale in the edge pixels on these filters, neural networks are used. Then the edge properties are improved at a suitable scale. Because of the above operation they can eliminate the structures that are blurred since those structures don`t have the specified scales. Text area detection was done on the enhanced images.

2.2. Texture-based Methods

This method is based on the fact that texts in images have the different textual properties than that from the background of the images. Basically this technique is based on wavelet, Gabor filters, spatial variance, FFT, etc. to find the textual properties of the text area.

Zhong et al. [16] proposed the method to locate the text area in the grayscale images based on local spatial variations. To find the spatial variance for pixels in the native neighborhood, they use the horizontal box of size 1×21. After that by using the canny edge detector, horizontal edges are determined. And then small edge components are merged into the larger lines. From the resultant edge image those text line are selected who is paired with lower and upper boundaries of edges in opposite direction. The limitation with this approach is that it can only detect the text that has the horizontal orientation compared to the background. The processing time of only 6.6 second was achieved when it was tested on the dataset containing the image of resolution 256×256 on a SPARC station 20. Zhong [27] presents another paper with similar approach for the I-frames and JPEG of compressed videos.

Park et al. [31] uses the same method for finding the vehicle license plate. For license plate localization they use the horizontal variation of text. The difference lies on the fact that they use the TDNN (time delay neural network) for the texture discriminator in the HSI color space. For the horizontal and vertical filters two TDNNs are used. To decide whether a window contains the license plate number, each HSI color value for a small window is passed through each neural network. At the end projection profile analysis was done to locate the bounding boxes for license plate. And this was done after merging the two filtered images.

Wu et al. [32,33] proposed a method based on multi-scale texture segmentation for the segmentation of an input image. By using the 2^{nd} –order Gaussian derivatives, potential text regions are selected. On each filtered image a non-linear transformation is applied. By using the output of the non-linear transformation local energy estimates are computed for each pixel and then clustering was done on the basis of the K-means algorithm.

Overall the process is called as the texture segmentation. After this, chip generation stage is initialized that contains the 5 steps a) stoke generation, b) stroke filtering c) stroke aggregation d) chip filtering e) chip extension. Both the stages – texture segmentation and chip generation should be done for multiple scales, so that text can be detected for the different range of size. The testing was done on the data set containing 48 images of the video frames, envelops, magazine, newspaper, etc. The processing time is almost 10 second when it is tested on the having 200 MHz Pentium pro processor with 128 MB memory. The problem with this method is that it will unable to detect the very small text. The system is able to detect more than 90% character if the character is larger than 10 pixels.

Sin et al. [34] proposed the algorithm in which they use the frequency features like number of pixels in the vertical and horizontal directions and Fourier spectrum to find the most likely text regions in the scenery image. On the fact that many text areas are on a rectangular background, and then by detection the edge rectangle search is performed, followed by the Hough transform.

Kim et al. [35] proposed the algorithm based on SVMs (Support Vector Machines) for examining the textual properties of texts in images. SVMs are a powerful technique and it works well with non-linear data and SVMs can also include a feature extractor within the architecture. Here the process of input feeding is same as that in Jain and Zhong [47]. After texture classification was done by using the SVMs, extraction of text lines was performed on the basis of profile analysis.

Chun et al. [36] used the mixture of neural network and FFT. To reduce the processing time, FFT is determined for overlapped line segments containing 1×64 pixels. 32 features were obtained for the output of each line segment. To differentiate the pixels, the feed-forward neural network with one hidden layer is used. The input vector for this neural network is 32 features that they obtained earlier. Labeling operations and noise elimination are performed on the image obtained as the output of neural network. The authors claim that there system can be used in the real time but actual processing time in any environment was not reported.

A machine learning based technique was proposed by Li et al. [37, 38] for tracking and localizing the text in the videos. To achieve this task, a small window of size 16×16 is applied to scan the images. The scanning was done for multiple times. Each window is then classified as the text and non-text based on the trained neural network. Here the features include the second and third order central moments of the decomposed sub-band image, and mean value for the same. To reduce the overall time, tracking process was implemented. Tracking process will also help to stabilize the localization results. To deal with the more complex motions, contour-based text stabilization method was used. For the experiment purpose various kinds of images are collected that contains movie credits, news, sports, videoconferences, and sports commercial. They include scene texts and captions having multiple fonts and sizes for the experiment. The procedure for the localization takes only 1 sec in Sun Ultra Workstation. For text detection, recall rate of 88% and precision rate of 62% was achieved.

T.Kumuda and L.Basavaraj [39] proposed an efficient algorithm for text detection and localization in natural images in 2013. The images contain valuable information and this information if extracted properly can be used for identity purpose, retrieval purpose or for indexing purpose. There are two stages in the proposed algorithm. In the first stage text regions are detected by using texture features and use of discriminative functions is made for filtering out the non -text regions. The second stage localization and merging of the detected text is performed. The proposed algorithm successfully detects and localize text of different styles, sizes, languages and orientations.

Traditional texture methods are computationally expensive. The large amount of the processing time is taken by the texture classification stage. This is because the texture-based filtering needs to scan the image to detect the text and confine the text regions.

2.3. Text Extraction in Compressed Domain

Now a day's most digital images and videos are stored, processed, and transmitted in a compressed form. There are some methods that have been presented recently like TIE methods that directly operate on images in JPEG or MPEG compressed formats. The advantage with these algorithms is that it require small amount of decoding scheme and

because of this property it is fast algorithm. Furthermore for the text detection, the methods like DCT coefficients and motion vectors in MPEG videos are useful.

Yeo and Liu [40] proposed a technique that can localize the caption text in a reduced resolution image. The benefit with the reduced resolution image is that it can easily be reconstructed from the compressed video by using the Ac and DC components from MPEG videos. Since the resolution is decreased by a factor of 16 or 64, resulting in poor text localization rate. Because the text area is localized based on a large interframe difference, therefore the rapid frame order changes can only be detected. The problem with this method is that it missed the scrolling titles and also it is limited by the assumption that texts only appear in predefined area of the video frame.

Gargi et al. [41] proposed an algorithm that has four stage approaches for TIE a) detection, b) localization, c) segmentation, d) recognition. They applied the text detection in a compressed domain, but they only use the number of intracoded block in P- and B-frames, without the I-frame of MPEG video sequences. This is based on the fact that when captions are appearing or disappearing then equivalent block are typically intracoded. This technique also gives the state of the art performance to abrupt scene changes or motion.

Zhong et al. [42] proposed the technique for localizing captions in JPEG images and I-frames of MPEG compressed videos. To capture textural properties, they use the DCT coefficient like periodicity and directionality of local image blocks. Then by using the connected component analysis and morphological operation, results are refined. As per the author it will take only 0.006 second to process the image of resolution 240×350 and has false alarm rate of 1.87% and recall rate of 99.17%. The problem with this algorithm is that the accurate localization outcomes could not be produced as all unit block is defined as non-text or text.

From this huge study of literature survey it is found that many different types of techniques were proposed by many researchers to detect and localize the text in scenery image or in video frames . These algorithms works well for detecting and localizing the

text but the problem with this algorithm is that they are producing a large number of false positive regions. False positive region in an image can be defined as non-text region incorrectly identified as text region. The gaps lie on the fact that none of the proposed algorithms are able to eliminate satisfactorily false positive regions. Even they had applied some filters to distinguish from text and non-text but they were not able to achieve a satisfactory result. Hence, it is required to fill these gaps. Therefore, a study has been proposed where an efficient algorithm was developed and implemented, which can eliminate almost all the false positive result from the output of any text detection algorithm.

Chapter 3 Problem Statement

Detection and localization of the text in the scenery image suffers from various types of problems due to complex background, low contrast, and lightning effects. These problems need to be removed before sending it to recognition phase of OCR system. Keeping these points in mind, the problem and objective of present study are discussed below.

3.1. Problem Statement

Text plays very important role in the scenery images or in video frames and understanding that text carries important information is the challenging task. Actually text detection is the area in which we try to find all the text regions in the scenery images or in video frames by finding precise boundaries of text lines through different approaches. Text detection phase is the important phase because the subsequent phases, segmentation, and text recognition all depends on the output of the text detection phase. The problems that making the detection and recognition of the text in the scenery image a difficult task are discussed below:

- Text can be appearing within objects or with the noisy background.
- There is lot of difference in the size, color, font, style, orientation, and alignment even inside the similar word.
- There is a chance that portion of the text excluded because of the camera and object movements.
- There may be complex movement in the scene.
- There may be a lot of lighting variation.
- There is a chance of deformation if text appears in the stretchy surface like cloth.
- Text may appear with the special arts or effects and with some images.

Till now a lot of research has been done in this area but there is no any robust solution that can work on all kind of scene text.

3.2. Thesis Objectives

Segmenting the text regions from the scenery images or in video frame containing text is the main objective. From the discussion done in chapter 1, it is clear that the overall accuracy of the photo OCR system will depend on the text detection phase. The better the performance of text detection phase the better will be the overall performance. There are many proposed methods to detect text in scenery image/video frames but the problem with this method is that they are producing the huge number of false positive image regions. Main objective of this thesis are:

- To analyze the output of MSER feature detector algorithm.
- To find which kind of structures falls in false positive region.
- To create the dataset of text region and non-text region based on the analysis.
- To apply the MSER feature detector on the test dataset.
- To reduce the false positive regions by applying some filters.
- To apply the text verifier (ANN- classifier) to verify the given region as text or non-text.

Chapter 4 Proposed Method

False positive regions are the problem with every text detection algorithm, so here new methodology that has been proposed with motive to reduce the false positive regions. Methodology aims to eliminate false positive regions in the result of MSER feature detector in less time.

4.1. Preliminaries

OCR: - OCR (optical character recognition) is the system through which we can covert the scanned document or document image clicked by camera into the editable text format.

ANN Classifier: - ANN is one the machine learning techniques which are inspired by biological neural system (the central nervous system in case animals) that are used to estimate functions that can depend on a huge amount of inputs and generally they are unknown. Basically ANN is presented in the form of interconnected neurons that can pass the message between each other. There is a weight on the each connection that can change them based on the training so that ANN can learn and can adapt to the new inputs. The classifier means here is that the output of ANN is discrete. The architecture of ANN classifier is shown in Figure 7.

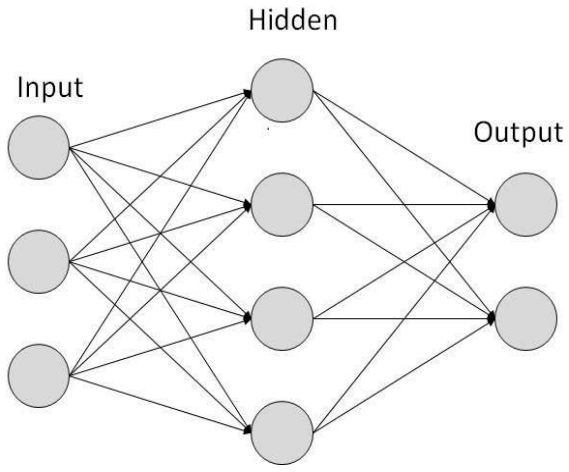


Figure 7: Architecture of ANN- classifier

4.2. Data Sets

A data set is collected that contains more than 90 scenery images with different variability in structure, color, blur, geometry and appearances. Let's name this dataset as dataset-1. Dataset-1 contains the images of street sign board, name plate, and advertisement board, shop board, and etc. with different lightning conditions. Street sign board is mainly collected from the Chandigarh. Direction board and name board are collected from the Thapar University. Some of the random scenery images of our dataset-1 are given in Figure 8. The resolutions of the images are vary from the 1823×1740 to 4224×2560.

Figure 8: Random Images of the Dataset 1

MSER feature detector is applied on all the samples images of the dataset-1 and from the output of the MSER, an analysis has been done to find out which kind of regions gives the false positive result. From the analysis it is found that background structures like tree leafs, tree branches, roofs, and windows of the building are mostly gives the false positive result. Another dataset is created from the above analysis which contains 100 images of non-text regions and 100 images for the text regions. Let's say this dataset as dataset-2. Sample images of the non-text regions in dataset-2 are shown in Figure 9 and the sample images of the text regions in dataset-2 is shown in Figure 10.

Figure 9: Dataset 2 Containing the Images of Non-text Regions

Figure 10: Dataset 2 Containing the Images of Text Regions

4.3. Implementation

Some times for the better performance hybrid approach are preferred. So for the better performance different methods are applied in the output of the MSER feature detector. The flow diagram to extract the text from the scenery image is given in Figure 11.

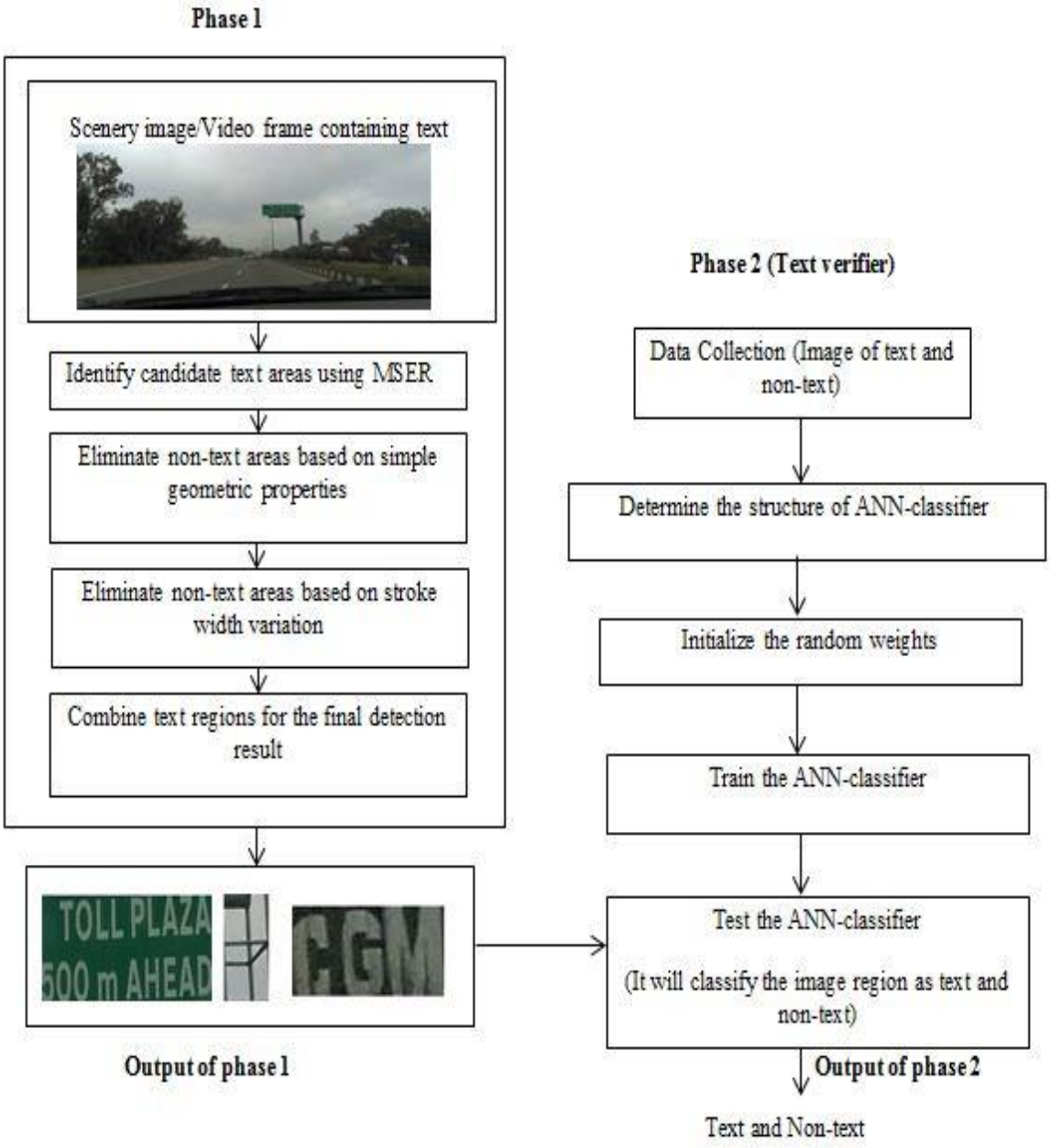


Figure 11: Flowchart of the proposed method

The flow diagram is divided into two phases, in the first phase MSER feature detector is applied and then several filters are applied to eliminate the non-text regions like elimination of non-text region based on simple geometric properties, and elimination of non-text region based on stroke width variationon the output of MSER feature detector After this the segmented text regions are fed into the ANN classifier which acts as the text verifier. The implementation of ANN classifier has been done to classify the input images as text or non-text. Given a 30-by-30 pixel segmented region, the ANN classifier will classify whether the window contains text regions or not. Here ANN classifier works as the binary classifier because it has two outputs (text/non-text). The ANN classifier is trained by using the dataset-2. The learning architecture of ANN-classifier is shown in Figure 11.

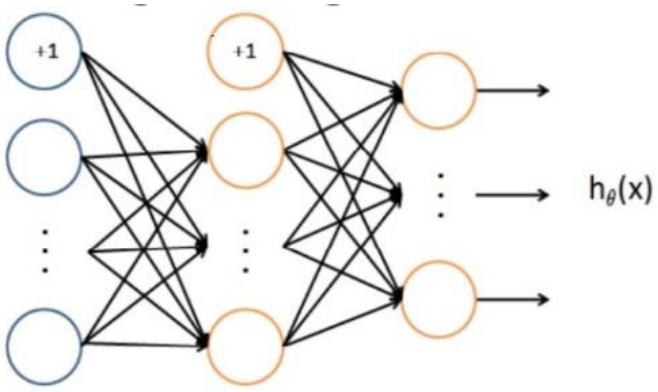


Figure 12: Learning Architecture of ANN-classifier

We preferred the ANN classifier because it is very efficient in learning the complex non-linear things. It has 3 layers – an input layer, a hidden layer and an output layer. Since the images are of size 30 × 30, which gives us 900 input layer units (not including the additional bias unit which always yields +1). The training data is loaded into the variables x and y and we randomly initialized the weight into the variable $\Theta^{(1)}$, $\Theta^{(2)}$ of the ANN. The datasets and the weight of the ANN classifier are stored into the separate mat file. The parameters have dimensions that are sized for an artificial neural network with 60 units in the second layer and 2 output units (belonging to the two classes' text/non-text).The 30 random grayscale images dataset 2 are shown in Figure 13.

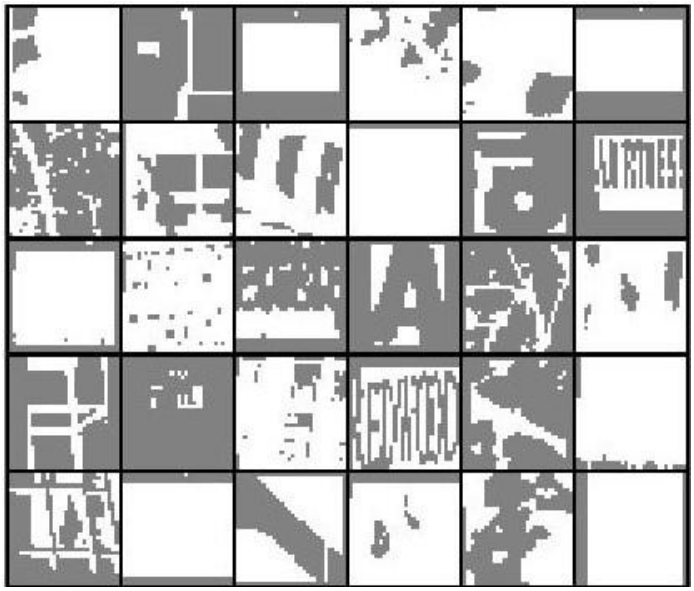


Figure 13: 30 Random Graysclae Images of Dataset 2

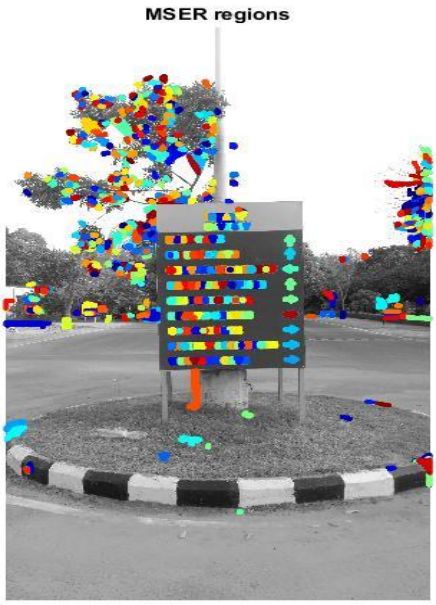
The sequence of the methods applied in scenery images to find the text region is given below:

Step 1 - Detection of candidate text region has been done by using the MSER feature detector. Figure 14 shows the a) original image and b) selected region after applying the MSER feature detector.

From the Figure 14, it is clear that many non-text regions are also get selected the a text region after applying the MSER feature detector.

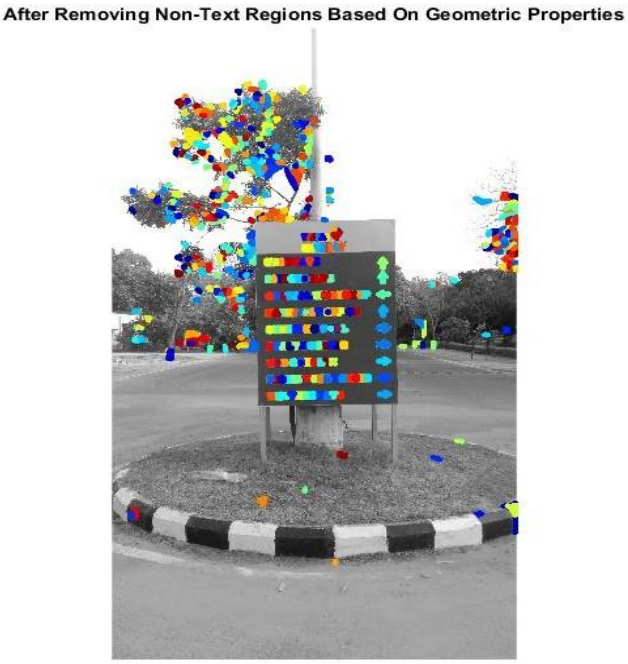
Step 2: Removal of non-text region on the basis of basic geometric properties.

- Different kind of geometric properties has been used to discriminate between the text and non-text regions like eccentricity, aspect ratio, extent, Euler number, and solidity based on simple threshold value.
- Figure 15shows the removal of non-text regions based on simple geometricproperties.



(a) **(b)**

Figure 14a) Original Image And b) Selected Region After Applying the MSER Feature Detector.



.

Figure 15: Removal of Non-text Regions Based on Simple Geometric Properties

Step 3: Removal of non-text regions based on the stroke width variation.

- The measurement of the stroke width variation is done on the basis of the width of the curve and lines that style up a character. There is a little variation in the stroke width of the text regions, whereas large variation in the stroke width of the non-text regions. To get the clear picture of how the stroke width has very small variation over other regions see the figure 16, and it is because the lines and the curves are generally have the equal widths.
- Figure 17 shows the effect after removing the non-text regions based on the stroke width variation.

:

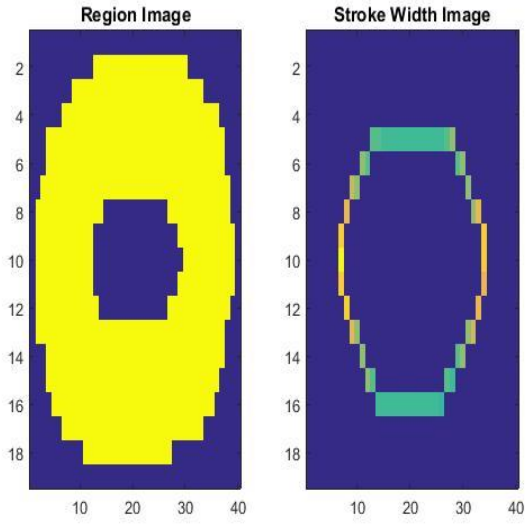


Figure 16: Small Variation of Stroke Width for the Character Components

Step 4: Merging of characters into word or text lines.

- Since the output of the step 3 is consist of individual characters. Neighboring text has been detected to merge the text regions into the word or lines. Figure 18 shows the result after applying the step 4 in terms of bounding boxes.
- Pairwise overlap ratio has been found and those bounding boxes whose overlap ratio is less than 2 are removed.

Figure 17: Effect after Removing the Non-text Regions Based on the Stroke Width Variation

Figure 18: Expanded Bounding Boxes Text

Step 5: ANN classifier- works as the verifier to verify the output of step 4 text regions, as the text and non-text. The different steps to implement the ANN classifier are given below:

1. Define the network architecture and prepare the training data set to train the neural network.
2. Regularized cost function has been used to overcome the problem of overfitting andunderfitting.

   The cost function is calculatedby using the equation 2 as given below:

$$J(\Theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[-y_k{}^{(i)}\log\left(\left(h_\Theta\left(x^{(i)}\right)\right)_k\right) - \left(1 - y_k{}^{(i)}\right)\log\left(1 - \left(h_\Theta\left(x^{(i)}\right)\right)_k\right)\right] + \frac{\lambda}{2m}\left[\sum_{j=1}^{60}\sum_{k=1}^{900}\left(\Theta_{j,k}{}^{(1)}\right)^2 + \sum_{j=1}^{60}\sum_{k=1}^{2}\left(\Theta_{j,k}{}^{(2)}\right)^2\right] \quad (2)$$

   Where, $J(\Theta) =$ Regularized cost function, m = number of pixels applied as the input in layer 1, j= number of units in second layer, K= number of output layer, $y_k{}^{(i)}$= value of output k in ith training example $x^{(i)}$ = ith training example, $\left(h_\ominus\left(x^{(i)}\right)\right)_k$= output value of K^{th} output unit, λ = Regularized parameter

3. Weights of ANN classifier have been initialized randomly and it is near to 0.
4. Then implementation of the forward propagation has been done to get the hypothesis. The various steps are given below:

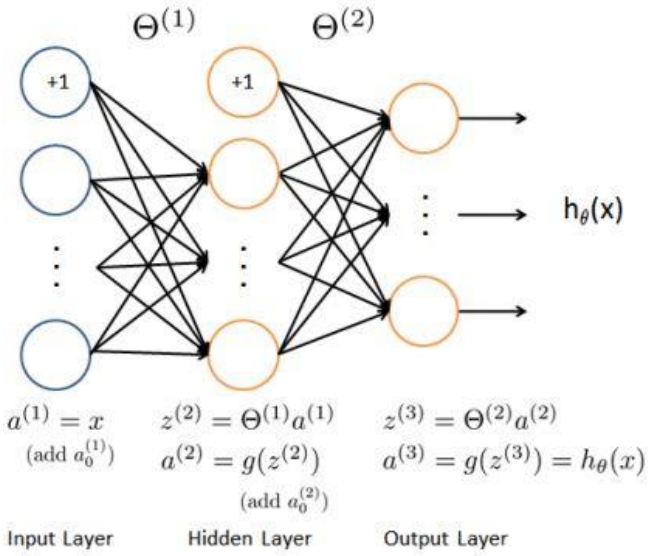


Figure 19: Forward Propagation

Given a training example ($x^{(t)}$,$y^{(t)}$), "forward pass" is done to compute all the activations throughout the network, including the output value of the hypothesis h_Θ(x).l is the number of layers,$a^{(l)}$=activation of layer l, Set the input layer's values ($a^{(1)}$) to the t[th] training example$x^{(t)}$. Then perform a feedforward pass (Figure 18), computing the activations ($z^{(2)}$,$a^{(2)}$,$z^{(3)}$,$a^{(3)}$) for layers 2 and 3 by using the equation 3. One need to add a+1 term to ensure that the vectors of activations for layers $a^{(1)}$and $a^{(2)}$also include the bias unit. In layer 1, input acts as the activation. Mathematically different steps are:

- $a^{(1)}$ = x
- $z^{(2)} = a^{(1)} \times \Theta^{(1)}$
- $sigmoid(z) = g(z) = \frac{1}{1+e^{-z}}$
- $a^{(2)} = g(z^{(2)})$
- $z^{(3)} = a^{(2)} \times \Theta^{(2)}$
- $a^{(3)} = g(z^{(3)})$

(3)

5. Implementation of the code to calculate the cost function has been done by using the equation number 2.

6. Implementation of the code for the back propagation to calculate the partial derivatives has been done. The various steps to perform this task is given below:

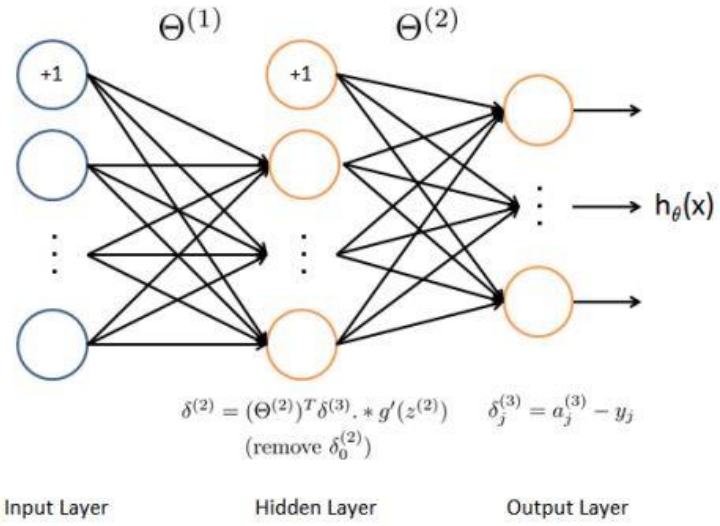


Figure 20: Backpropagation Updates

6.1. For each output unit k in layer 3 (the output layer), the error is calculated by equation 4.

$$\delta_k^{(3)} = (a_k^{(3)} - y_k) \tag{4}$$

Where $a_k^{(3)}$= activation of unit k in layer 3

Where $y_k \in \{1,2\}$ indicates whether the current training example belongs to class k (y_k= 1), or if it belongs to a different class (y_k= 2).

6.2. For the hidden layer l = 2, the error is calculated by equation 5.

$$\delta^{(2)} = \left(\Theta^{(2)}\right)^T \delta^{(3)} .* g'(z^{(2)}) \tag{5}$$

6.3. Accumulation of the gradient from is done by using the following equation 6

$$\Delta^{(l)} = \Delta^{(l)} + \delta^{(l+1)}\left(a^{(l)}\right)^T \tag{6}$$

6.4. The (unregularized) gradient for the neural network cost function is obtained by dividing the accumulated gradients by $\frac{1}{m}$ as given in the equation 7.

$$\frac{\partial}{\partial \Theta_{ij}^{(l)}} J(\Theta) = D_{ij}^{(l)} = \frac{1}{m} \Delta_{ij}^{(l)} \tag{7}$$

7. Then gradient checking is used to compare the partial derivatives and numerical estimation of gradient of cost function.

Gradient checking is performed on the parameters, it can be done by imagine "unrolling"the parameters$\Theta^{(1)}$, $\Theta^{(2)}$into a long vector θ. By doing so, one can think ofthe cost function being J (θ) instead and use the following gradient checkingprocedure.Function $f_i(\theta)$ that purportedly computes $\frac{\partial}{\partial \theta_i} J(\theta)$; it is better to check if $f_i$is outputting correct derivative values.

$$\left.\begin{aligned} \text{Let } \Theta^{(i+)} &= \Theta + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \varepsilon \\ \vdots \\ 0 \end{bmatrix} \text{and } \Theta^{(i-)} = \Theta - \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \varepsilon \\ \vdots \\ 0 \end{bmatrix} \end{aligned}\right\} \tag{8}$$

One can find the $\Theta^{(i+)}$and $\Theta^{(i-)}$ by using the equation 8. So, $\Theta^{(i+)}$is the same as θ, except its ith element has been incremented byε. Similarly, $\Theta^{(i-)}$is the corresponding vector with the ith element decreased byε.One can now numerically verify $f_i(\theta)$'s correctness by checking, for eachi, by using the equation 9:

$$f_i(\Theta) \approx \frac{J\left(\Theta^{(i+)}\right) - J\left(\Theta^{(i-)}\right)}{2\varepsilon} \tag{9}$$

8. Advanced optimization technique has been used with back propagation to minimize the cost function.

There are total of 161 text regions detected by the phase 1 of the proposed method form the dataset-3. All the 161 images have been fed into the ANN classifier and it gives the output for each text regions as the text (prediction is 1) and non-text (prediction is 2) as shown in Figure 21.

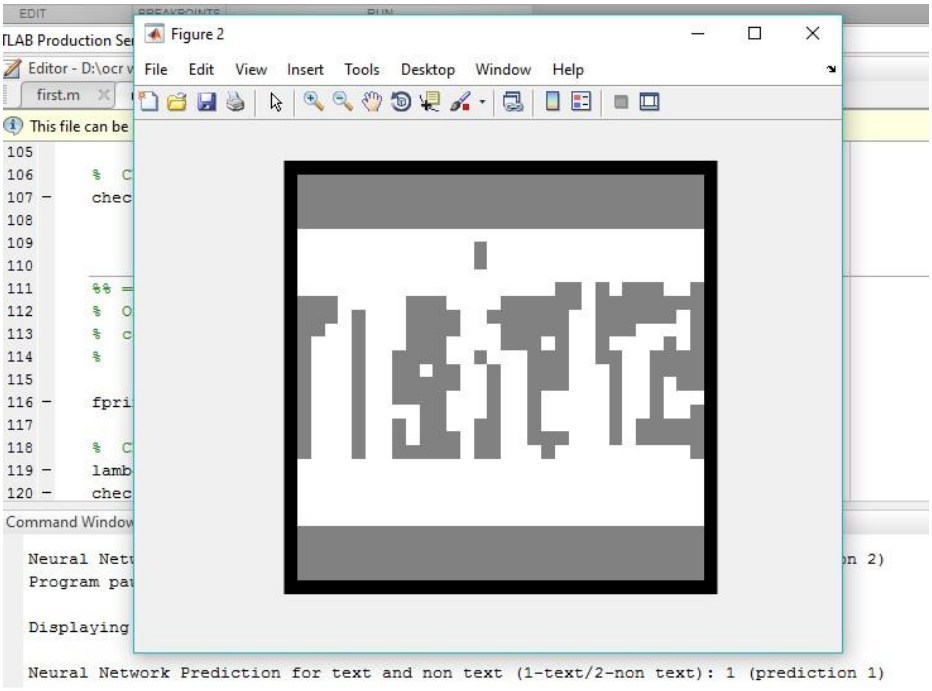


Figure 21: Output of Trained ANN-classifier For Each Image Region as Text or Non-text

Chapter 5 Experimental Results

This chapter includes analysis that is performed when proposed method is applied to dataset in terms of objectives that are achieved. Comparative analysis shows that proposed method provides better results in term of precision, recall, and f-score. Results have been described in tabular form.

5.1. Implementation

All the codes are implemented in Matlab 2015a. Algorithm of Phase 1 is applied on the dataset-3 that contains more than 25 scenery images. One of image from dataset-3 is shown in Figure 22, andafter applying the phase 1 algorithm, the output is shown in Figure 23.Results in terms of True positive, false positive and false negative for each image of dataset-3 is shown in table 1. All the terms are defined below:

- True positive (TP) is taken as text area correctly identified as text area.
- False positive (FP) as non-text area incorrectly identified as text area.
- True Negative (TN) as non-text area identified as non-text area.
- False negative (FN) as text area incorrectly identified as non-text area.

Segmented text regions, which is the output of the phase 1 is fed into the ANN classifier. The output of one of the segmented text region from the Figure 23 is given in Figure 24. Results in terms of true positive, false positive and false negative for each image of dataset-3 is shown in Table 2.

Figure 22: Original Image

Figure 23: Segmented Text Regions (Output of Phase 1 Algorithm)

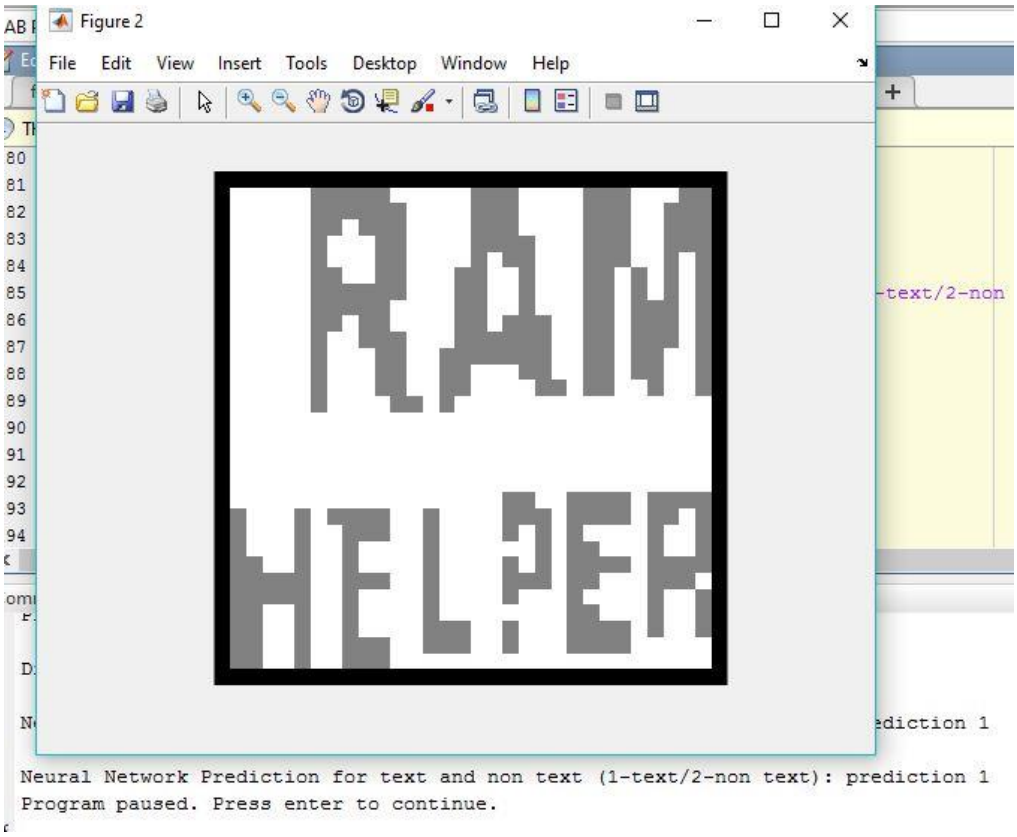


Figure 24: Output of Trained ANN-classifier for one of the ImageRegion from Figure 22 as Text or Non-text

There is a difference between phase 2 (ANN classifier) and the sliding window method. Because in sliding window method several scan should be perform with different aspect ratio to find the text in the image, and the result of this method is individual characters. In this case, the ANN classifier classifies the data based on group of the text and non-text. Since the ANN classifier is trained by the data set containing text regions (group of text) and non-text. From the Table 1, it can be analyzed that some images have lots of false positive result, like image number 22 has the highest false positive result and it is because of the fact that image number 22 has lots of complex structure in the background.

As shown in Table 2, it can be analyzed that after applying the text verifier the false positive result is almost eliminated from the output. Hence it will improve the overall accuracy of the OCR system.

Table 1: Output of MSER Feature Detector after Applying Several Filters on Each Image

Image	True positive	False Positive	False negative	True negative
1	2	1	0	0
2	4	8	0	0
3	1	2	0	0
4	4	8	0	0
5	1	7	0	0
6	3	4	0	0
7	2	2	0	0
8	1	3	0	0
9	15	0	3	0
10	3	6	0	0
11	5	0	0	0
12	3	0	0	0
13	1	0	0	0
14	1	0	0	0
15	1	5	0	0
16	1	5	0	0
17	1	8	0	0
18	9	5	0	0
19	1	0	1	0
20	2	1	2	0
21	1	0	0	0
22	6	16	0	0
23	1	0	0	0
24	3	1	0	0
25	2	6	1	0
26	4	5	6	0

Table 2: Output of Each Image afterApplying the Text Verifier (ANN Classifier)

Image	True positive	False Positive	False negative	True Negative
1	2	0	0	1
2	3	0	1	8
3	1	0	0	2
4	4	0	0	8
5	1	1	0	7
6	3	0	0	4
7	2	0	0	2
8	1	0	0	3
9	13	0	5	0
10	3	0	0	6
11	5	0	0	0
12	3	0	0	0
13	1	0	0	0
14	1	0	0	0
15	1	1	0	4
16	1	1	0	4
17	1	0	0	8
18	9	0	0	5
19	1	0	1	0
20	2	0	2	1
21	1	0	0	0
22	6	3	0	13
23	1	0	0	0
24	2	0	1	1
25	2	1	1	5
26	3	2	7	3

5.2. Results

The experimental result of algorithm is evaluated in terms of precision rate, recall rate, and f-score values. Precision rate is defined as all the text in image where we predict the text, what fraction of it is actually a text. Recall rate is defined as all the text that actually is text, what fraction us correctly detected as having text. Mathematically

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$F = \frac{1}{\frac{\alpha}{precision} + \frac{1-\alpha}{Recall}}$$

To combine the precision and recall standard f-measure has been used. The parameter α control the relative weight α has been set to 0.5 so that it gives the equal weight to precision and recall. If α =0.75 (recall weighted 0.75, precision weighted 0.25), and hence unbalanced weight, the α=0.75 setting indeed produced higher recall rate.The phase 1 gives the precision rate of 0.46 and the recall rate of 0.85. After integrating theANN classifier (text verifier-phase 2) the precision rate isincreased, and has the value of 0.89, and recall rate of 0.80and it is because the fact that text area which is not extracted in phase 1 cannot be extracted in phase 2. The results are represented in the tabular form as follows in the table3. For the performance comparison, the results of ICDAR dataset are taken from the [6]. The special dataset in the table 3 refers to non-standard datasets.

Table 3: Performance Comparison of Text Detection Algorithm

Methods	Dataset	Recall	Precision	F-score
Proposed algo	Dataset-3	0.80	0.89	0.84
Gllavata et.al	Special	0.80	0.83	0.81
Alex Chen	ICDAR	0.60	0.60	0.58
Qiang Zhu	ICDAR	0.40	0.33	0.33
Jisoo Kim	ICDAR	0.28	0.22	0.22
Hinnerk Becker	ICDAR	0.62	0.67	0.58
Nobuo Ezaki	ICDAR	0.36	0.18	0.22
Ashida	ICDAR	0.46	0.55	0.50
HWDavid	ICDAR	0.46	0.44	0.45
Wolf	ICDAR	0.44	0.30	0.35
Todoran	ICDAR	0.18	0.19	0.18
Epshtein	ICDAR	0.60	0.73	0.66

Chapter 6 Conclusion and Future Scope

This chapter includes the brief description of what the problem was and measures that are achieved and scope for further research has been stated clearly under the future scope section.

6.1. Conclusion

The present study of text detection system has been created based on a hybrid approach containing two phases. The first phase is connected component approach (MSER) with several filters and the second phase is the machine learning approach (ANN-classifier). The proposed algorithm in scenery images containing text is applied and it is found thatthe phase 1 gives the precision rate of 0.45 and the recall rate of 0.84. After integrating the ANN classifier (text verifier-phase 2) the precision rate is increased, and has the value of 0.89, and recall rate of 0.80. To improve the performance of the system, the phase 2 (text verifier) can be integrated in any method whose output gives huge false positive result. It is demonstrated that by finding which type of structure in scenery images gives the false positive result and by making that type of structure to train the ANN-classifier gives the best result. The text verifier is less computationally expensive then sliding window method because instead of performing several scans in the text regions and then classifying it as text and non-text, this method will do it in only one step**.** Our proposed methodology achieves the top precision and recall value compared to other methods. Even though the training of the ANN classifier has been done with dataset containing only 200 images, but it gives the best result.

6.2. Future Scope

The approach presented can be improvised or extended in the following ways:

- Some preprocessing like increasing the contrast and sharpness of the image could be performedto improve the performance of MSER feature detector.
- For the classification purpose some more powerful algorithms like SVM (Support Vector Machine), kernels, etc. could be applied in phase 2.

- Further improvement in the performance can be achieved by adding more training set images on phase-2.
- New features can be added to improve the accuracy of text detection.
- For the better performance, Phase 2 can be integrated in any method whose output gives huge false positive result.

REFERENCES

[1] Jung, Keechul, KwangIn Kim, and Anil K. Jain. "Text information extraction in images and video: a survey." *Pattern recognition* 37, no. 5 (2004): 977-997.

[2] Chen, Datong, and JuergenLuettin. *A survey of text detection and recognition in images and videos*. No. EPFL-REPORT-82656.IDIAP, 2000.

[3]Zhang, Jing, and RangacharKasturi. "Extraction of text objects in video documents: Recent progress." In *2008 The Eighth IAPR International Workshop on Document Analysis Systems*. 2008.

[4]Doermann, David, Jian Liang, and Huiping Li. "Progress in camera-based document image analysis." In *Document Analysis and Recognition, 2003.Proceedings. Seventh International Conference on*, pp. 606-616. IEEE, 2003.

[5] Jain, Anil K., and Yao Chen. "Address block location using color and texture analysis." *CVGIP: Image Understanding* 60, no. 2 (1994): 179-190.

[6] Chen, Xiangrong, and Alan L. Yuille. "Detecting and reading text in natural scenes."In *Computer Vision and Pattern Recognition, 2004.CVPR 2004.Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II-366.IEEE, 2004.

[7] Lienhart, Rainer, and Axel Wernicke. "Localizing and segmenting text in images and videos." *IEEE Transactions on circuits and systems for video technology* 12, no. 4 (2002): 256-268.

[8] Yangxing, L. I. U., and Takeshi IKENAGA. "A contour-based robust algorithm for text detection in color images." *IEICE transactions on information and systems* 89, no. 3 (2006): 1221-1230.

[9] Gllavata, Julinda, Ralph Ewerth, and Bernd Freisleben. "Text detection in images based on unsupervised classification of high-frequency wavelet coefficients." In *Pattern Recognition, 2004.ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1, pp. 425-428. IEEE, 2004.

[10] Jain, Anil K., and Bin Yu. "Automatic text location in images and video frames."In *Pattern Recognition, 1998.Proceedings. Fourteenth International Conference on*, vol. 2, pp. 1497-1499. IEEE, 1998.

[11] Kim, Hae-Kwang. "Efficient automatic text location method and content-based indexing and structuring of video database." *Journal of Visual Communication and Image Representation* 7, no. 4 (1996): 336-344.

[12] Li, Huiping, David Doermann, and OmidKia. "Automatic text detection and tracking in digital video." *IEEE transactions on image processing* 9, no. 1 (2000): 147-156.

[13] Ye, Qixiang, Qingming Huang, Wen Gao, and Debin Zhao. "Fast and robust text detection in images and video frames." *Image and Vision Computing* 23, no. 6 (2005): 565-576.

[14] Jain, Anil K., and Sushil K. Bhattacharjee. "Address block location on envelopes using Gabor filters." *Pattern Recognition* 25, no. 12 (1992): 1459-1477.

[15] Jain, Anil K., and FarshidFarrokhnia. "Unsupervised texture segmentation using Gabor filters." *Pattern recognition* 24, no. 12 (1991): 1167-1186.

[16] Zhong, Yu, KalleKaru, and Anil K. Jain. "Locating text in complex color images." In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1, pp. 146-149. IEEE, 1995.

[17] Ohya, Jun, Akio Shio, and Shigeru Akamatsu. "Recognizing characters in scene images." *IEEE transactions on pattern analysis and machine intelligence* 16, no. 2 (1994): 214-220.

[18] Lee, Chung-Mong, and AtreyiKankanhalli. "Automatic extraction of characters in complex scene images." *International Journal of Pattern Recognition and Artificial Intelligence* 9, no. 01 (1995): 67-82.

[19]Shim, Jae-Chang, ChitraDorai, and Ruud Bolle."Automatic text extraction from video for content-based annotation and retrieval."In *Pattern Recognition,*

1998.Proceedings. Fourteenth International Conference on, vol. 1, pp. 618-620. IEEE, 1998..

[20] Jain, Anil K., and Bin Yu. "Automatic text location in images and video frames."In *Pattern Recognition, 1998.Proceedings. Fourteenth International Conference on*, vol. 2, pp. 1497-1499. IEEE, 1998..

[21] Messelodi, Stefano, and Carla Maria Modena. "Automatic identification and skew estimation of text lines in real scene images." *Pattern Recognition* 32, no. 5 (1999): 791-810.

[22] Jung, Keechul, and JungHyun Han. "Texture-based text location for video indexing."In *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 449-454.Springer Berlin Heidelberg, 2000.

[23] Hase, Hiroyuki, Toshiyuki Shinokawa, Masaaki Yoneda, and Ching Y. Suen. "Character string extraction from color documents." *Pattern Recognition* 34, no. 7 (2001): 1349-1365.

[24]. Weinman, Jerod, Allen Hanson, and Andrew McCallum. "Sign detection in natural images with conditional random fields." In *Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*, pp. 549-558. IEEE, 2004.

[25]. Epshtein, Boris, EyalOfek, and Yonatan Wexler. "Detecting text in natural scenes with stroke width transform." In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2963-2970. IEEE, 2010..

[26].Huang, Shen, Wang, and Gao, C., , January. "Text detection and recognition in natural scene images."In *Estimation, Detection and Information Fusion (ICEDIF), 2015 International Conference on* (pp. 44-49). IEEE, 2015

[27] Smith, Michael A., and Takeo Kanade. *Video skimming for quick browsing based on audio and image characterization*.School of Computer Science, Carnegie Mellon University, 1995.

[28] Sato, Toshio, Takeo Kanade, Ellen K. Hughes, and Michael A. Smith. "Video OCR for digital news archive."In *Content-Based Access of Image and Video Database, 1998.Proceedings., 1998 IEEE International Workshop on*, pp. 52-60. IEEE, 1998.

[29] Hasan, Yassin MY, and LinaKaram. "Morphological text extraction from images." *IEEE Transactions on Image Processing* 9, no. 11 (2000): 1978-1983..

[30] Chen, Datong, Kim Shearer, and HervéBourlard. "Text enhancement with asymmetric filter for video OCR."In *Image Analysis and Processing, 2001.Proceedings. 11th International Conference on*, pp. 192-197. IEEE, 2001..

[31] Park, Sung Han, Kwang In Kim, Keechul Jung, and Hyung Jin Kim. "Locating car license plates using neural networks." *Electronics Letters* 35, no. 17 (1999): 1475-1477.

[32] Wu, Victor, RaghavanManmatha, and Edward M. Riseman. "Textfinder: An automatic system to detect and recognize text in images." *IEEE Transactions on pattern analysis and machine intelligence* 21, no. 11 (1999): 1224-1229.

[33] Wu, Manmatha, and R. Riseman, "Finding Text in Images."*Proc. of ACM International Conference on Digital Libraries*, 1997, pp. 1-10.

[34] Sin, Bong-Kee, Seon-Kyu Kim, and Beom-Joon Cho. "Locating characters in scene images using frequency features." In *Pattern Recognition, 2002.Proceedings. 16th International Conference on*, vol. 3, pp. 489-492. IEEE, 2002.

[35] Shin, C. S., K. I. Kim, M. H. Park, and Hang Joon Kim. "Support vector machine-based text detection in digital video." In *Neural networks for signal processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop*, vol. 2, pp. 634-641.IEEE, 2000.

[36] Chun, Byung Tae, YounglaeBae, and Tai-Yun Kim. "Automatic text extraction in digital videos using FFT and neural network."In *Fuzzy Systems Conference Proceedings, 1999.FUZZ-IEEE'99. 1999 IEEE International*, vol. 2, pp. 1112-1115. IEEE, 1999.

[37Li, Huiping, David Doermann, and OmidKia. "Automatic text detection and tracking in digital video." *IEEE transactions on image processing* 9, no. 1 (2000): 147-156.

[38] Li, Huiping, and David Doermann. "A video text detection system based on automated training."In *Pattern Recognition, 2000.Proceedings. 15th International Conference on*, vol. 2, pp. 223-226. IEEE, 2000.

[39] Kumuda, T., and L. Basavaraj. "Detection and localization of text from natural scene images using texture features." In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1-4.IEEE, 2015.

[40] Yeo, Boon-Lock, and Bede Liu. "Visual content highlighting via automatic extraction of embedded captions on MPEG compressed video." In *Electronic Imaging: Science & Technology*, pp. 38-47. International Society for Optics and Photonics, 1996.

[41] Antani, S., U. Gargi, D. Crandall, T. Gandhi, Ryan Keener, and R. Kasturi. "A system for automatic text detection in video."In *ICDAR'99*, p. 29.IEEE, 1999.

[42] Zhong, Yu, Hongjiang Zhang, and Anil K. Jain. "Automatic caption localization in compressed video." *IEEE transactions on pattern analysis and machine intelligence* 22, no. 4 (2000): 385-392.

List of Publications

Communicated:

1. Animesh Sharma, Rajiv Kumar, A hybrid approach to reduce the false positive result of text detection in scenery images,Journal of Visual Languages and Computing.

Video Link