RECOGNITION OF MULTI-TYPE AND MULTI-ORIENTED TEXT IN VIDEOS

SANGHEETA ROY

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

RECOGNITION OF MULTI-TYPE AND MULTI-ORIENTED TEXT IN VIDEOS

SANGHEETA ROY

THESIS SUBMITTED IN FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Sangheeta Roy

Matric No: WHA140025

Name of Degree: Doctor of Philosophy

Title of Project Paper/Research Report/Dissertation/Thesis ("RECOGNITION OF MULTI-TYPE AND MULTI-ORIENTED TEXT IN VIDEOS"):

Field of Study: Image Processing, Text Recognition

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

RECOGNITION OF MULTI-TYPE AND MULTI-ORIENTED TEXT IN VIDEOS

ABSTRACT

Text inscribed in video plays an important role to understand the semantic essence of the content in several real-time application, such as video events indexing and retrieval, license plate recognition, automatic navigation, and surveillance applications. Since video suffers from multi-text type, multi-oriented text, low resolution, complex background, thus achieving accurate recognition results is challenging and interesting. In general text appearance and background in video differs according to application and problems. Therefore, in this thesis, a new method has been proposed based on texts and its background to classify the video type, which results in the video of particular text type. To enhance the video images from the effect of Laplacian operation, fractional Poisson model has been introduced for removing noise introduced by Laplacian operation in the video. A multimodal approach is explored for detecting words in complex video images, such as sports, Marathon video images, etc. which can cope with the causes of background and foreground variations. Then detected words are used for keyword spotting in the video to retrieve the video frames efficiently. Since keyword spotting does not involve semantic information to retrieve the video events, a new classification algorithm has been proposed based on tampered and context features to classify the caption and scene text types which facilitates recognition to achieve good recognition rate. To recognize the text in video images, Bayesian classifier-based method has been investigated for binarization to use available OCR. However, the primary focus of this approach limits to horizontal English texts. Therefore, Hidden Markov Model-based recognition method which works without binarization has been proposed for recognizing the text of multiple scripts. The proposed methods are evaluated over standard datasets and our own datasets using standard evaluation metrics. Furthermore, the proposed methods are compared with existing recent methods to show that proposed methods outperform the existing methods in terms of quality and quantity measures.

Keywords: Multi-type Text; Multi-oriented Text; Text Recognition.

RECOGNITION OF MULTI-TYPE AND MULTI-ORIENTED TEXT IN VIDEOS

ABSTRAK

Teks tertulis dalam video memainkan peranan yang penting untuk memahami intipati semantik kandungan dalam beberapa aplikasi masa sebenar, seperti pengindeksan dan dapatan semula acara video, navigasi automatik, pengecaman plat lesen dan aplikasi pengawasan. Oleh kerana video mengalami jenis teks majmuk, teks orientasi majmuk, resolusi rendah, latar belakang kompleks, mencapai keputusan pengecaman yang tepat menjadi cabaran yang besar dan menarik. Secara umumnya, pertunjukan teks dan latar belakang dalam video berbeza mengikut aplikasi dan pelbagai masalah. Oleh itu, dalam tesis ini, kami mencadangkan satu kaedah baru berdasarkan teks-teks dan latar belakangnya untuk mengklasifikasikan jenis video, yang menyebabkan video jenis teks yang tertentu. Untuk meningkatkan imej video dari kesan operasi Laplace, kami memperkenalkan model Poisson pecahan untuk mengeluarkan bunyi yang diperkenalkan oleh operasi Lapalcian dalam video. Kami meneroka pendekatan multimodal untuk mengesan perkataan dalam imej video yang kompleks, seperti sukan, imej video Marathon dan lain-lain yang boleh menampung dengan punca-punca variasi latar belakang dan latar depan. Seterusnya, kata-kata yang dikesan akan digunakan untuk mengesan kata kunci dalam video untuk mendapat semula bingkai video dengan cekap. Oleh kerana mengesan kata kunci tidak melibatkan maklumat semantik untuk mendapatkan semula peristiwa video, kami mencadangkan klasifikasi algoritma baru berdasarkan ciri-ciri yang diganggu dan konteks untuk mengklasifikasikan kapsyen dan tempat kejadian jenis teks yang memudahkan pengecaman untuk mencapai kadar pengecaman yang tinggi. Untuk mengecam teks dalam imej-imej video, kami mencadangkan kaedah berasaskan pengelas Bayesian supaya binarisasi boleh menggunakan OCR yang sedia ada. Walau bagaimanapun, fokus utama kaedah ini mengehadkan teks bahasa Inggeris mendatar. Oleh itu, kami mencadangkan Kaedah pengecaman berdasarkan Hidden Markov Model tanpa binarisasi untuk mengecaman teks berbilang skrip.

Kaedah yang dicadangkan dinaksir berdasarkan piawai dataset dan dataset kami sendiri yang menggunakan metrik penaksiran piawai. Tambahan pula, kaedah yang dicadangkan dibandingkan dengan kaedah yang sedia ada pada tahun-tahun kebelakangan ini agar membuktikan bahawa kaedah yang dicadangkan lebih berkesan daripada kaedah yang sedia ada dari segi kaedah kualitatif dan kuantitatif.

Kata kunci: Teks berbilang jenis; Teks berorientasikan pelbagai; Pengiktirafan teks.

ACKNOWLEDGEMENTS

I would like to thank everyone who inspired and helped me during my doctoral study.

First of all, I would like to express my heartiest appreciation and sincere gratitude to my supervisors Dr. Palaiahnakote Shivakumara, Dr. Ainuddin Wahid Abdul Wahab, and Dr. Umapada Pal. Their support and guidance from the basic to the concluding level enabled me to develop an understanding of the subject.

My sincere thanks also go to Aniruddha Sinha, project leader of Tata Consultancy Services who provided me an opportunity and valuable support for continuing my doctoral study.

I am indebted to my present and ex-colleagues of Multimedia lab of the University of Malaya (UM) for a very pleasant and friendly working environment. I sincerely appreciate Vijeta with whom I had many technical discussions and I am thankful for her help during my stay in UM. I also want to thank Pavani, Sarahbukhari, Ishwarya, Saqib, and others for their valuable support which helped me stay focused on my graduate study.

I am also grateful to all the administrative personnel for their valuable support.

Lastly, and most important, my profound gratitude to my family. Words cannot express my feelings nor my thanks that how grateful I am to my dearest mother, father, elder sister, elder brothers, and parents-in-law. Your blessings, love, and care have been my constant motivation towards fulfilling my dreams and aspirations in life. Special gratitude goes to my elder brother and mentor Partha Roy for unflagging support, guidance, and without whom, I would never have the courage to embark on this journey. And my profound gratitude to my loving husband for always standing by my side and keeping faith in me. To them, I dedicate this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

2D	:	Two-dimensional
3D	:	Three-dimensional
AB	:	After Binarization
AIV	:	Adjusted Intensity Value
APT	:	Average Processing Time
BB	:	Before Binarization
BDI	:	Born Digital Images
BNDCG	:	Binary Normalized Discounted Cumulative Gain
BoW	:	Bag of Words
CBIR	:	Content Based Image Retrieval
CC	:	Connected Component
CLAHE	:	Contrast-Limited Adaptive Histogram Equalization
CNN	:	Convolutional Neural Network
CRF	:	Conditional Random Field
CW	:	Total Number of Correct Word
DCT	:	Discrete Cosine Transform
DWT	:	Dynamic Wavelet Transform
F	:	F-measure
FCN	:	Fully Convolutional Network
GVF	:	Gradient Vector Flow
HE	:	Histogram Equalization
HMM	:	Hidden Markov Model
HOG	:	Histogram Oriented Gradients
IDWT	:	Inverse Dynamic Wavelet Transform

LBP	:	Local Binary Pattern
LM	:	Language Model
LSTM	:	Long short-term memory
MAP	:	Maximum a Posteriori
MLP	:	Multi-Layer Perceptron
MRF	:	Markov Random Field
MSER	:	Maximal Stable External Region
MSE	:	Mean Standard Error
MW	:	Total Number of Matched Word
NDCG	:	Normal Discounted Cumulative Gain
NLP	:	Natural Language Processing
NN	:	Neural Network
NZC	:	Non-Zero Coefficient
NW	:	Total Number of Word
OCR	:	Optical Character Recognizer
P	:	Precision
PCA	:	Principal Component Analysis
PHOG	:	Pyramidal Histogram of Oriented Gradient
pLSA	:	Probabilistic Latent Sematic Analysis
PSNR	:	Peak signal-to-noise ratio
QA	:	Quality Assessment
R	:	Recall
RGB	:	Red Green Blue
RNN	:	Recurrent Neural Network
RR	:	Recognition Rate
SIFT	:	Scale Invariant Feature Transform

SSIM	:	Structural Similarity
SVM	:	Support Vector Machine
SVT	:	Street View Data
SWT	:	Stroke Width Transform
TCR	:	Text Candidate Region
SIFT	:	Scale Invariant Feature Transform
SURF	:	Speeded Up Robust Features
SVM	:	Support Vector Machine
SWT	:	Stroke Wavelet Transform
WGF	:	Wavelet-Gradient Fusion
WHO	:	World Health Organization
ZC	:	Zero Coefficient

CHAPTER 1: INTRODUCTION

1.1 Introduction

In the last few years, with the recent progress of science and technology, especially the evolution of mobile, use of video and images for daily activities of human being increases drastically, thereby resulting in huge demand in information retrieval field (N. Sharma, Pal, & Blumenstein, 2012). According to official statistics, almost 300 hours of videos are uploaded in YouTube per minute, nearly 5 billion videos are watched in every single day, and more than half of YouTube views come from mobile devices. From another popular social networking site Facebook, it is declared that an average of 8 billion daily video views from 500 million users, which was 4 billion views in April 2016. The wide usage of multimedia (image, video) in shape of communication, educational and entertaining, needs robust annotation or recognition of text for indexing and retrieving the text accurately in minimal time. This is because extracting relevant information from huge database efficiently is a hard task for content-based image retrieval methods (Doermann, Liang, & Li, 2003). The main reason for poor results of the content-based image retrieval methods is that difference between the high-level features and low level (Lyu, Song, & Cai, 2005; Jing Zhang & Kasturi, 2008). Due to this gap, the methods are not adequate to interpret the meaning of the content in the video or images. To fill such gap in image/video, text detection and recognition become popular. The text identification and recognition help in obtaining the essence which is much related to content. Therefore, accurate and efficient text detection and recognition by overcoming different challenges posed by different applications have become challenging and interesting.

Detection and recognition of text is not a new issue in document analysis community. It can be seen that many Optical Character Recognizer (OCR) engines have been used for the different script in the literatures (U. Pal & B. Chaudhuri, 2001). However, the video and scene text cannot be feed directly to OCR as these OCR were developed for plain document images where homogenous background and high contrast exist. In case of video or scene images, one can expect multiple adverse factors such as variant font style, different font size, complex background, contrast, oriented text, multi-type text, etc. and the effect of uneven illumination (Lyu et al., 2005; Q. Ye & Doermann, 2015). The work discussed the issues with existing OCR engines and challenges of video and scene text recognition in detail in subsequent sections. For recognizing text, Optical Character Recognition (OCR) system is the way of converting the image into corresponding readable text. In other words, the aim of the system was to convert a given image to digitized image such that system can understand the content of the image. OCR translates from one script to another script and retrieves the documents from the large database automatically through tags (Moghaddam & Cheriet, 2010; Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010; P. Ye & Doermann, 2013). OCR has been evolving over many years to fit into the new types of applications starting from scanner to mobile which are the different medium of image captured devices. In the following sub-sections, these will be described in more details.

1.1.1 OCR for Scanned Document Image

In the beginning, the target was to digitize the documents of plain background scanned by a scanner device. Therefore, the steps of general OCR for such document image can be seen in Figure 1.1, where one can see the significant steps, preprocessing, line segmentation, character segmentation and text recognition (Ahmad et al., 2016; Mithe, Indalkar, & Divekar, 2013). It is noted that while scanning the document image, there were two common causes, such as noise introduced by the device and tilt in the image due to improper document placement over the scanner.

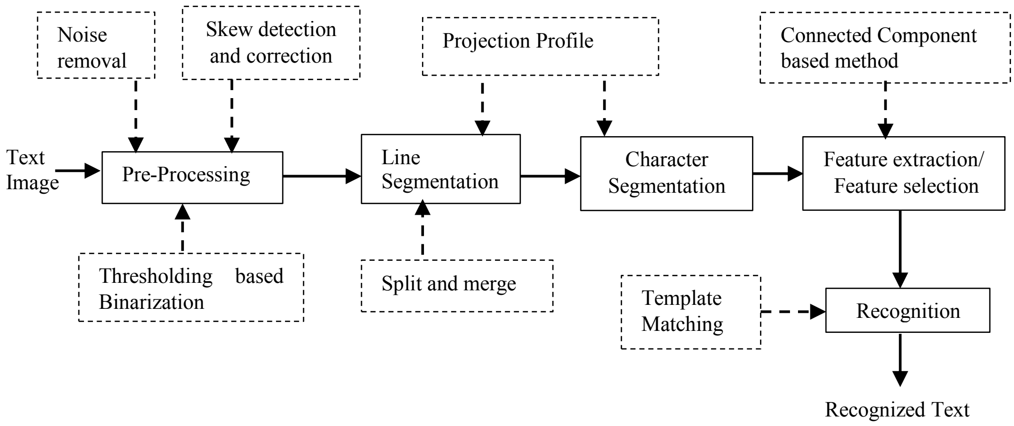
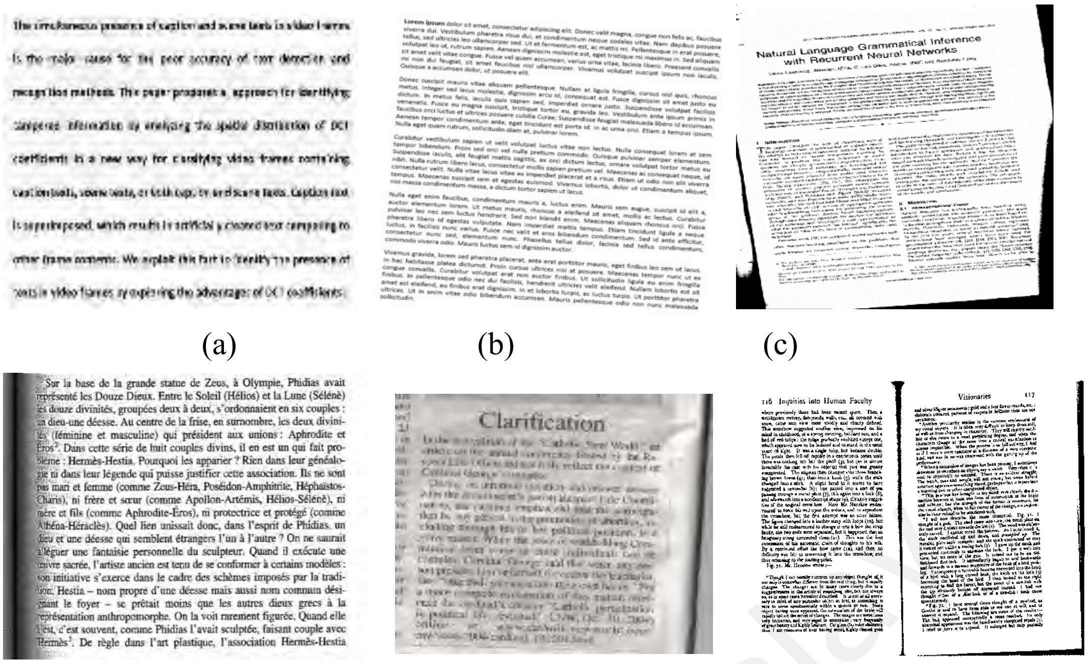


Figure 1.1: OCR for plain document image.

These two causes are the primary issue for not achieving good recognition results from the document images (Farahmand, Sarrafzadeh, & Shanbehzadeh, 2017). To prevent the effect of the above causes, the methods were developed for noise removal, skew detection and correction which were called pre-processing methods. Since the target of the OCR development was plain images, the methods were developed for binarizing the image where the text is separated from the background. As a result, the system gets a binary image for the given input image. For the purpose of text line and character segmentation, most of the methods explore projection profiles by taking advantage of the binary form of the image. The segmented character can be matched with the predefined templates for recognition. Figure 1.2 shows some examples of the scanned image. A proper scanned document can be seen in Figure 1.2 (a). Figure 1.2 (b, c) depicts skewed images, and Figure 1.2 (d, e, f) represents noise in scanned book and article.



The simultaneous presence of caption and scene texts in video frames is the major cause for the poor accuracy of text detection and recognition methods. This paper proposes a approach for identifying tampered information by analyzing the spatial distribution of DCT coefficients in a new way for classifying video frames containing caption texts, scene texts, or both caption and scene texts. Caption text is superimposed, which results in artificially created text comparing to other frame contents. We exploit this fact to identify the presence of texts in video frames by exploring the advantage of DCT coefficients.

(a)

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec velit magna, congue non felis ac, faucibus viverra dui. Vestibulum pharetra risus dui, et condimentum neque sodales vitae. Nam dapibus posuere tellus, sed ultricies leo ullamcorper sed. Ut et fermentum est, ac mattis mi. Pellentesque in erat posuere, volutpat leo ut, rutrum sapien. Aenean dignissim molestie est, eget tristique mi maximus in. Sed aliquam sit amet velit vitae congue. Fusce vel quam accumsan, varius urna vitae, lacinia libero. Praesent convallis mi non dui feugiat, sit amet faucibus nisl ullamcorper. Vivamus volutpat suscipit ipsum non iaculis. Quisque a accumsan dolor, ut posuere elit.

Donec suscipit mauris vitae aliquam pellentesque. Nullam at ligula fringilla, cursus nisl quis, rhoncus metus. Integer sed lacus molestie, dignissim arcu id, consequat est. Fusce dignissim sit amet justo eu dictum. In metus felis, iaculis quis sapien sed, imperdiet ornare justo. Suspendisse volutpat facilisis venenatis. Fusce eu magna suscipit, tristique tortor eu, gravida leo. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Suspendisse feugiat malesuada libero id accumsan. Aenean tempor condimentum ante, eget tincidunt est porta id. In ac urna orci. Etiam a tempus ipsum. Nulla eget quam rutrum, sollicitudin diam at, pulvinar lorem.

Curabitur vestibulum sapien ut velit volutpat luctus vitae non lectus. Nulla consequat lorem at sem tempor bibendum. Proin sed orci vel nulla pretium commodo. Quisque pulvinar semper elementum. Suspendisse iaculis, elit feugiat mattis sagittis, ex orci dictum lectus, ornare volutpat tortor metus eu nibh. Nulla rutrum libero lacus, consectetur mollis sapien pretium vel. Maecenas ac consequat neque, id consectetur velit. Nulla vitae lacus vitae ex imperdiet placerat et a risus. Etiam ut odio non elit viverra tempus. Maecenas suscipit sem et egestas euismod. Vivamus lobortis, dolor ut condimentum aliquet, nisi massa condimentum massa, a dictum tortor sapien ut lacus.

Nulla eget enim faucibus, condimentum mauris a, luctus enim. Mauris sem augue, suscipit id elit a, auctor elementum lorem. Ut metus mauris, rhoncus a eleifend sit amet, mollis ac lectus. Curabitur pulvinar leo nec sem luctus hendrerit. Sed non blandit enim. Maecenas aliquam rhoncus orci. Fusce pharetra libero id egestas vulputate. Nam imperdiet mattis tempus. Etiam tincidunt ligula a neque luctus, in facilisis nunc varius. Fusce nec velit et eros bibendum condimentum. Sed id ante efficitur, consectetur nunc sed, elementum nunc. Phasellus tellus dolor, lacinia sed tellus condimentum, commodo viverra odio. Mauris luctus sem id dignissim auctor.

Vivamus gravida, lorem sed pharetra placerat, ante erat porttitor mauris, eget finibus leo sem ut lacus. In hac habitasse platea dictumst. Proin cursus ultrices nisi at posuere. Maecenas tempor nunc ut ex finibus. In pellentesque odio nec dui facilisis, hendrerit ultricies velit eleifend. Nullam lobortis est sit amet est eleifend, eu finibus erat dignissim. In et lobortis turpis, ac luctus turpis. Ut porttitor pharetra ultrices. Ut in enim vitae odio bibendum accumsan. Mauris pellentesque odio non nunc malesuada sollicitudin.

(b)

Natural Language Grammatical Inference with Recurrent Neural Networks

(c)

Sur la base de la grande statue de Zeus, à Olympie, Phidias avait représenté les Douze Dieux. Entre le Soleil (Hélios) et la Lune (Sélénè) les douze divinités, groupées deux à deux, s'ordonnaient en six couples : un dieu-une déesse. Au centre de la frise, en surnombre, les deux divinités (féminine et masculine) qui président aux unions : Aphrodite et Éros[2]. Dans cette série de huit couples divins, il en est un qui fait problème : Hermès-Hestia. Pourquoi les apparier ? Rien dans leur généalogie ni dans leur légende qui puisse justifier cette association. Ils ne sont pas mari et femme (comme Zeus-Héra, Poséidon-Amphitrite, Héphaistos-Charis), ni frère et sœur (comme Apollon-Artémis, Hélios-Sélénè), ni mère et fils (comme Aphrodite-Éros), ni protectrice et protégé (comme Athéna-Héraclès). Quel lien unissait donc, dans l'esprit de Phidias, un dieu et une déesse qui semblent étrangers l'un à l'autre ? On ne saurait alléguer une fantaisie personnelle du sculpteur. Quand il exécute une œuvre sacrée, l'artiste ancien est tenu de se conformer à certains modèles : son initiative s'exerce dans le cadre des schèmes imposés par la tradition. Hestia – nom propre d'une déesse mais aussi nom commun désignant le foyer – se prêtait moins que les autres dieux grecs à la représentation anthropomorphe. On la voit rarement figurée. Quand elle l'est, c'est souvent, comme Phidias l'avait sculptée, faisant couple avec Hermès[3]. De règle dans l'art plastique, l'association Hermès-Hestia

(d)

Clarification

(e)

116 Inquiries into Human Faculty

Visionaries 117

(f)

Figure 1.2: Examples of scanned documents where (a) Text in proper scanned image, (b) and (c) Skewed book and article, (d), (e), and (f) Noisy documents.

1.1.2 OCR for Camera-based Image

However, it is realized that scanner is expensive and it is not feasible for scanning a large number of images (Doermann et al., 2003). In addition, color is limited to the scanner (Doermann et al., 2003; P. Ye & Doermann, 2013). As a result, due to the advancement of new technologies, camera devices are introduced to replace scanner for capturing images. With this revolution, the OCR development is also extended to recognize complex documents such as degraded, distorted, historical, etc. in contrast to plain document image obtained by the scanner. Though camera devices are portable and provide flexibility in capturing images in terms of resolution, colors, etc., it introduces distortion such as low contrast, blur, perspective, multiple views, etc. while capturing images in contrast to scanner device. In order to reduce the effect of such issues, the new preprocessing methods are developed for enhancing the captured camera images (Jagannathan & Jawahar, 2005) as shown in Figure 1.3. For recognition, the method follows the same steps of scanned OCR. In Figure 1.4, some samples of camera-based images are shown where ‘a’ shows the proper camera based image, (b) and (c) are old and degraded documents, (d) is blurred and, (e) and (f) are perspective distorted images.

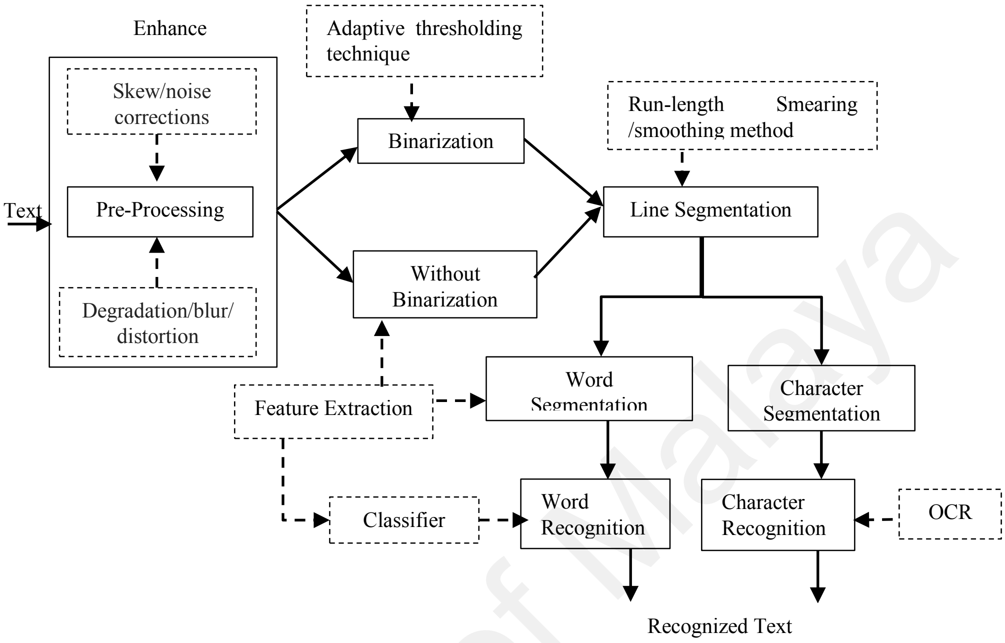
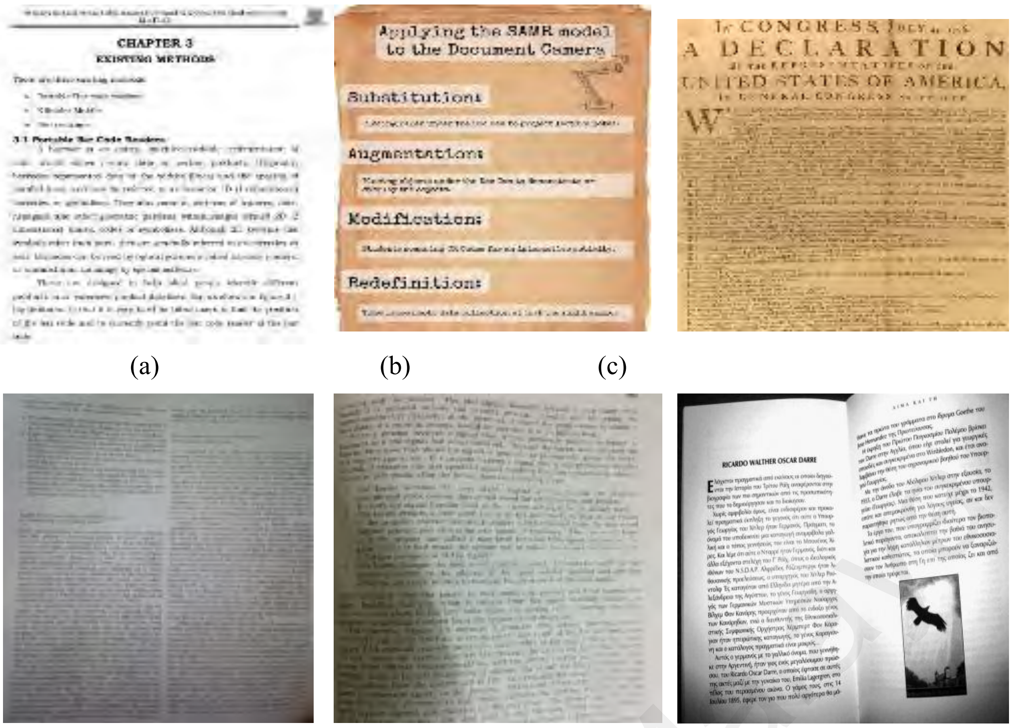


Figure 1.3: OCR for degraded, historical, blurred and distorted documents.



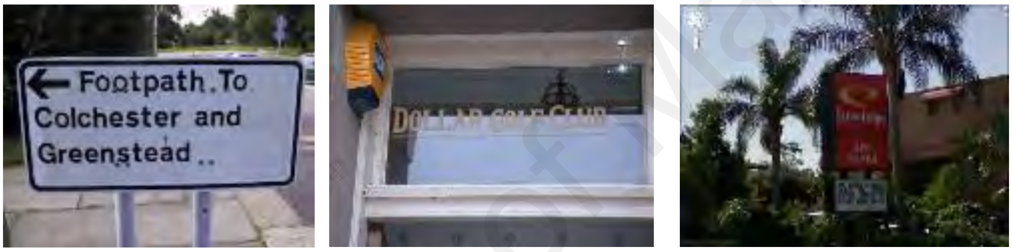
(a) (b) (c)

(d) (e) (f)

Figure 1.4: Examples of camera based images where (a) Text in camera based image, (b) and (c) Degraded and old historical article, (d) Blurred article, (e) and (f) Distorted and skewed article.

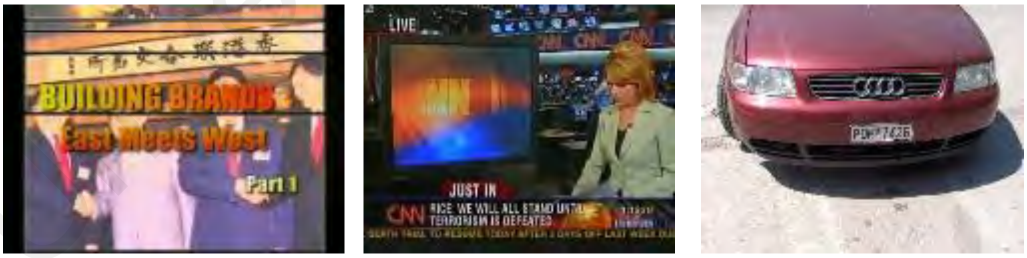
1.1.3 OCR for Natural Scene and Video Image

As a number of new applications mentioned in Section 1.1 emerge such as retrieving natural scene images, video images which contain text information for labeling images and video, license plate recognition to trace the vehicle, and assisting tourist person to translate one script to another script, the requirement and complexity of understanding images increase drastically (Q. Ye & Doermann, 2015; H. Zhang, Zhao, Song, & Guo, 2013). As a result, the image can suffer from the complex background, contrast, font, font size, orientation variations and distortion compared to clean document images. It is evident from the sample images shown in Figure 1.5 for natural scene images where one can see the complexity of the images for text detection and recognition. Similarly, Figure 1.6 shows sample images for video where low contrast text with different applications can be seen. Figure 1.7 shows the sequence of temporal information for the same text. Therefore, to develop OCR for such images, the main steps are seen in Figure 1.8. Since the images are complex compared to document images as discussed in the above, Figure 1.8 shows the step called text detection which has been introduced for separation of text region from the non-text region. In case of video, since it provides temporal information, it can be explored in detecting and recognizing text for improving the performance. It is a major advantage in the video compared to natural scene images. Text detection helps in separating text as foreground and non-text as background resulting in a binary image. Once the method gets the binary image, the steps that discussed for camera-based images can be applied for the scene/video text recognition with modifications (Yingying Zhu, Yao, & Bai, 2016).



(a) (b) (c)

Figure 1.5: Different types of scene text where (a) and (b) Horizontal text in street map and organization name, (c) Non-horizontal text in shop name.



(a) (b) (c)

Figure 1.6: Different types of video text, where (a), (b) Horizontal caption and scene text in news video, (c) Non- horizontal text in license plate.

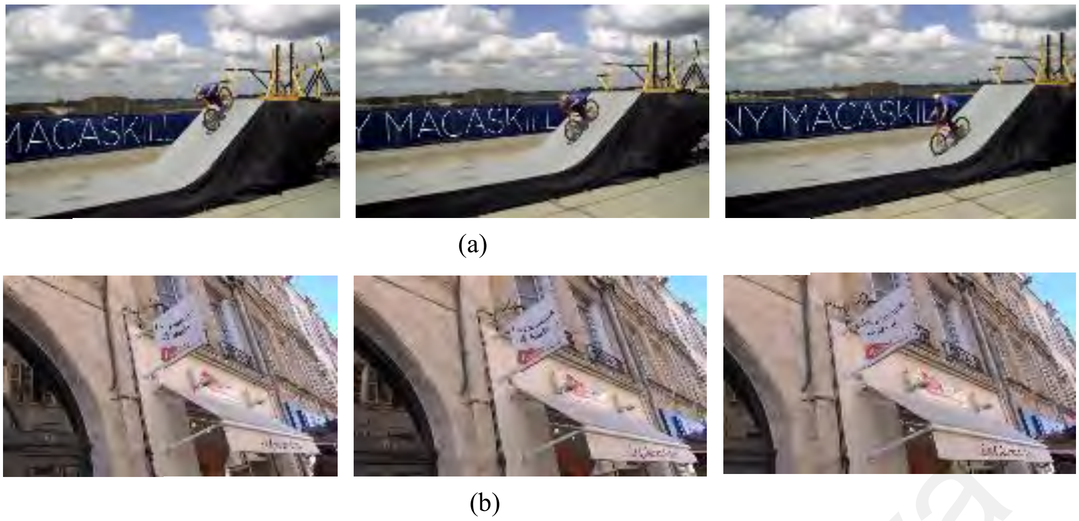


Figure 1.7: Different types of video text with temporal frames where, (a) Temporal frames containing distorted text of sports video, (b) Temporal frames containing curved text of street video.

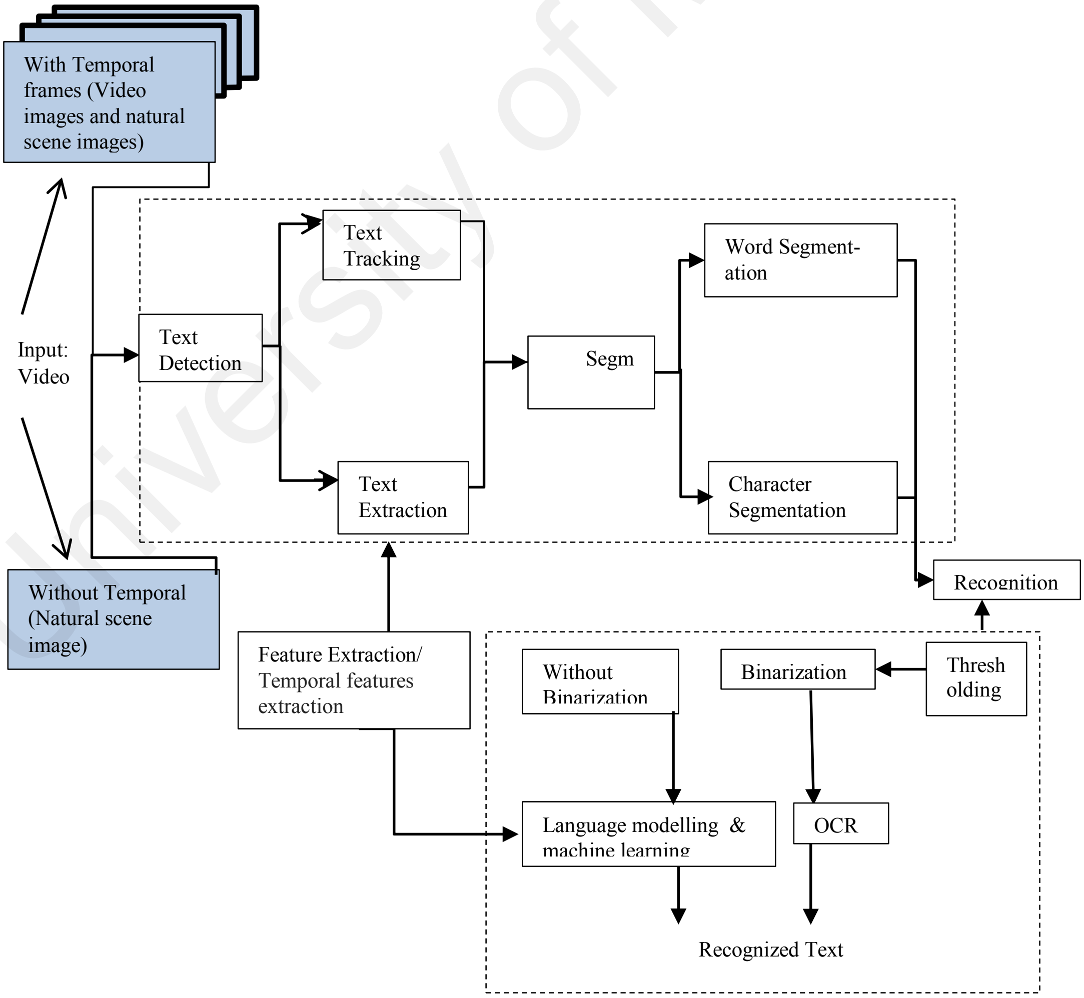

With Temporal frames (Video images and natural scene images)
Input: Video
Without Temporal (Natural scene image)
Text Detection
Text Tracking
Text Extraction
Segm
Word Segment-ation
Character Segmentation
Recognition
Feature Extraction/ Temporal features extraction
Without Binarization
Binarization
Thresh olding
Language modelling & machine learning
OCR
Recognized Text

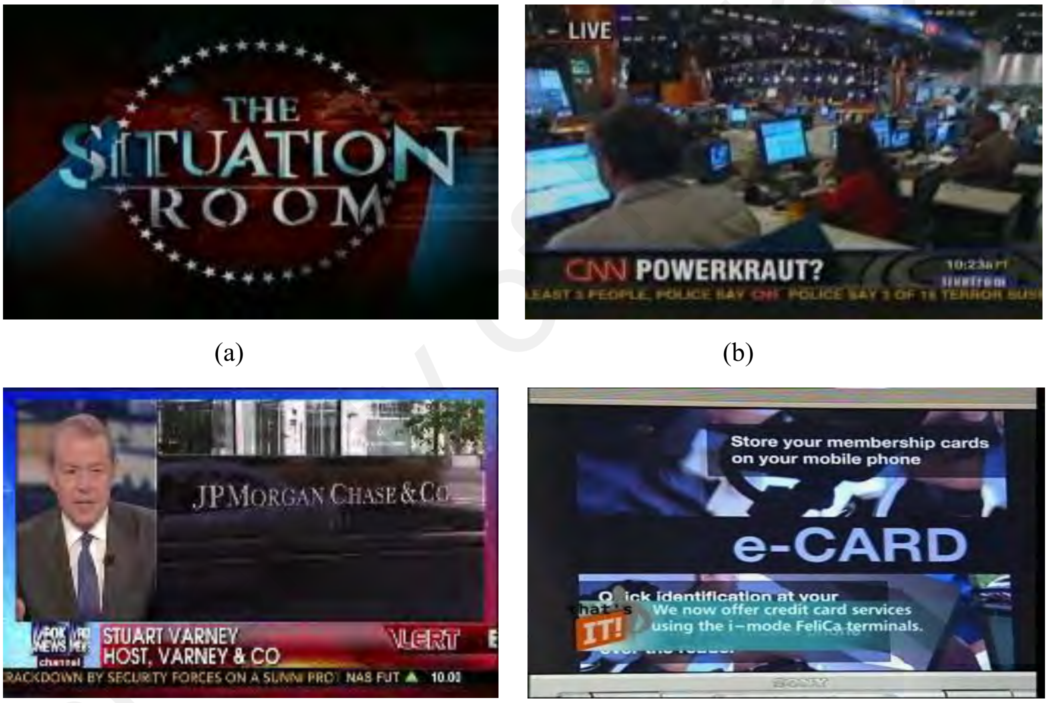
Figure 1.8: Natural Scene/ Video based OCR.

1.2 Motivation

As discussed in the above section and observed from Figure 1.5, Figure 1.6, and Figure 1.7 that most of the existing method focused on applications where simple text exists such as horizontal text, a particular type of text, etc. However, in reality, there are new applications, such as tracing the person in sports, marathon video and different types of the video where one can expect multi-type text, multi-fonts, multi-font size, arbitrarily oriented text apart from contrast and background variations. These challenges make text detection and recognition more complex and interesting (Grafmüller & Beyerer, 2013; J.-G. Park & Kim, 2013; Risnumawan, Shivakumara, Chan, & Tan, 2014). These factors motivated to divide the whole problem into several sub-problems, namely multi-type video text recognition, multi-oriented text recognition and multi-type-oriented text recognition which will be discussed in subsequent sections.

1.2.1 Multi-Type Video Text Recognition

For video understanding through text information, there are many methods developed in the recent time (D. Chen & Odobez, 2005; Saidane & Garcia, 2007; Z. Zhang & Wang, 2013b). However, the performances of these methods are not consistent and not satisfactory. The main reason is that there are two texts-types in every single frame in the video, namely, scene and caption text. Scene text appears naturally in images whereas caption text is manually edited. The two types of texts differ in their nature and characteristics. Caption text has good contrast, quality, clarity, uniform color, uniform text size, font, and very often it is in the horizontal direction and displayed at the bottom of the video since it is edited. Besides, scene text has the following properties, namely, low resolution, complex background, variations in font or font size, non-uniform illumination effect, blur effect due to text or camera movements, etc. and might appear anywhere in the frame. For example, videos of news especially sport which contains courts of different sports with captions (e.g., occurrence of a sports event, score summary) and scene texts (e.g., player name or bib number). Figure 1.9 depicts some examples of multi-type video text frames where (a) and (b) contains scene and caption type of news event, respectively. In Figure 1.9 (c) and (d) are the example of video frames of sports news where caption and scene text, both appear in the same frame. Such multi-type texts create much complex and challenging issues in case of text detection and recognition. Therefore, achieving better results for video with a single method is hard compared to those images having only one type of text, such as natural scene images and document images.


THE SITUATION ROOM
LIVE
CNN POWERKRAUT?
JPMORGAN CHASE & CO
STUART VARNEY
HOST, VARNEY & CO
ALERT
Store your membership cards on your mobile phone
e-CARD
We now offer credit card services using the i-mode FeliCa terminals.
IT!

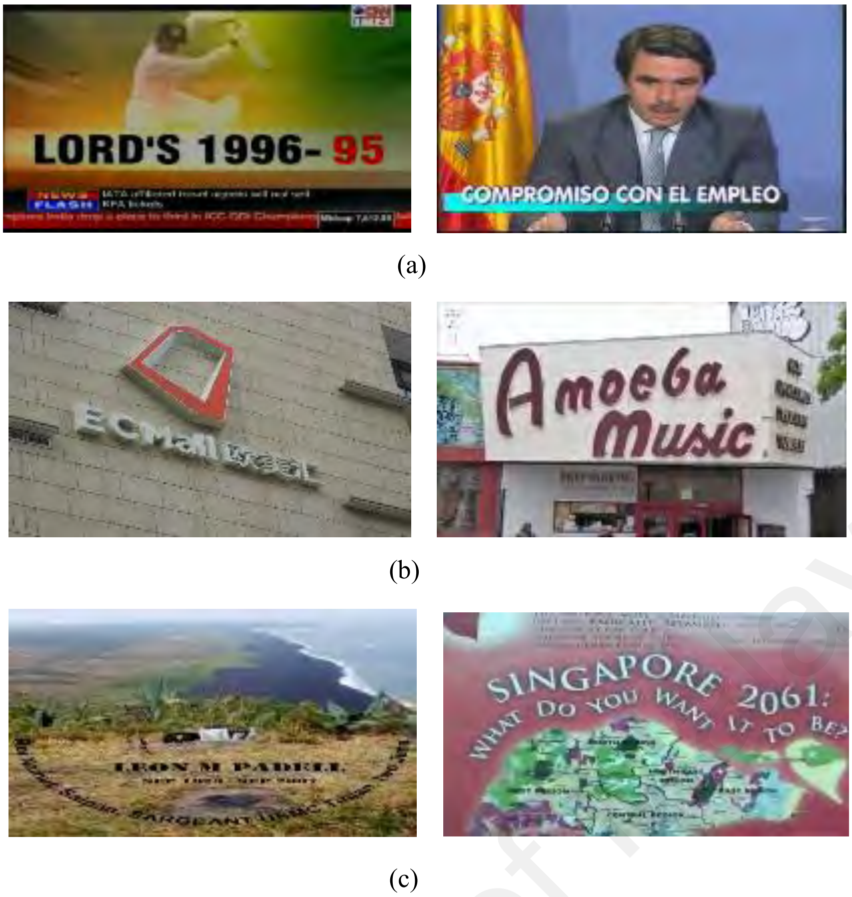
(a) (b)

(c) (d)

Figure 1.9: Examples of multi-type video text frame where (a) contains scene text type, (b) contains caption type, and (c) and (d) contain both caption and scene types in news video.

1.2.2 Multi-Oriented Video Text Recognition

As pointed out in the previous section, it is expected that the text of any orientation will be in video/ scene images according to applications. Therefore, recognizing arbitrarily oriented text is difficult because of extracting features without including background information unlike horizontal text. For instance, displayed text in shop name, banner, and sports including player name become more stylish which makes detection and recognition more complex. Therefore, the challenges in text detection and recognition mainly comes from three aspects: (1) Diversity of character appearances in word or text line, (2) Complexity of cluttered backgrounds and (3) Arbitrary orientation of text in the images. The existing document OCR and camera-based OCR cannot handle non-horizontal and curvy text. Most of the existing video-based OCR (Epshtein et al., 2010; K. I. Kim, Jung, & Kim, 2003; K. Wang, Babenko, & Belongie, 2011a) concentrate on recognizing horizontal or slightly non-horizontal texts, which will be discussed in next section of literature survey. Obviously, the pre-requisite of being horizontal extremely shrink the applicability of those methods in scenarios where images are taken casually under less controlled conditions as discussed in the beginning of the section. Thereby, multi-oriented video text recognition without any restriction on the background is more challenging. Specifically, curved text causes the recognition method to suffer more. Figure 1.10 depicts some examples of different orientated video text. Horizontal text can be seen in news video as shown in Figure 1.10 (a). Figure 1.10 (b) shows the non-horizontal text of building name and shop name due to unconstraint capturing and graphics style, respectively. Curvy text appeared in poster and banner are shown in Figure 1.10 (c), where varying oriented characters are seen in single text line and word.



(a)

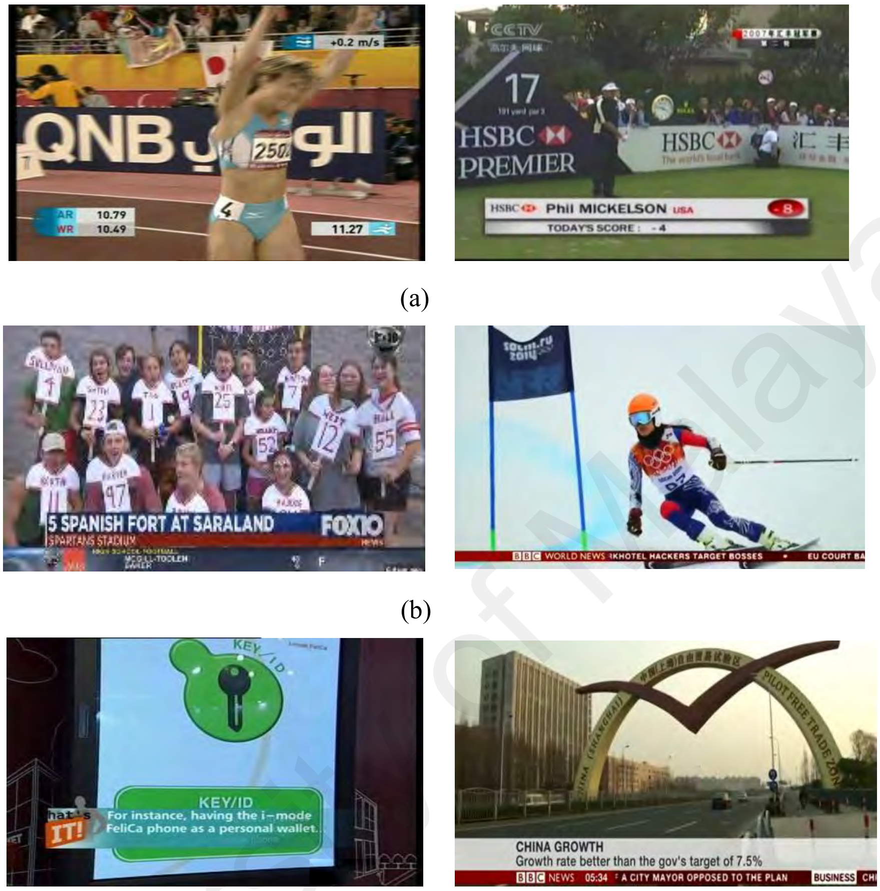
(b)

(c)

Figure 1.10: Examples of multi-oriented video text frame where (a) Horizontal text in news video, (b) Non-Horizontal text in shop and building and (c) Curvy text in poster.

1.2.3 Multi-Type Oriented Video Text Recognition

There are chances of images containing both multi-type with different orientation text especially sports and marathon video where can see caption text, scene text of different orientations. Figure 1.11 portrays some examples of multi-type video text having different orientation, especially for scene text. In these examples, all caption text appear in a horizontal way but scene text in the different orientation. In Figure 1.11 (a) and (b), scene text of person's bib number, placard and background board are seen in the horizontal and non-horizontal way for sports video whereas curvy scene text can be seen in wallboard and road in Figure 1.11 (c) for news video. These images are much more complicated than multi-type text images and multi-oriented text images. There are methods which address the issue of both type and orientation text individually. However, when the methods are performed on the images where both the complexities exist, the method reports inconsistent results and poor results. Therefore, this is still considered as an unsolved problem for text detection and recognition.



QNB
10.79
10.49
11.27
+0.2 m/s
CCTV
17
HSBC
PREMIER
Phil MICKELSON USA
TODAY'S SCORE : -4
5 SPANISH FORT AT SARALAND
SPARTANS STADIUM
FOX10
sochi.ru 2014
BBC WORLD NEWS HOTEL HACKERS TARGET BOSSES
KEY/ID
For instance, having the i-mode FeliCa phone as a personal wallet.
PILOT FREE TRADE ZONE
CHINA GROWTH
Growth rate better than the gov's target of 7.5%
BBC NEWS 05:34 A CITY MAYOR OPPOSED TO THE PLAN
BUSINESS

(a)

(b)

(c)

Figure 1.11: Examples of multi-type oriented video text frame where (a) and (b) Caption text in horizontal and scene text in non-horizontal way for sports video, and (c) Caption text in horizontal and scene text in curvy way in news video.

1.3 Challenges

Based on the above discussion, the challenges are listed and summarized for detecting and recognizing text in video and natural scene images.

- Figure 1.12 and Figure 1.13 show that video can contain multiple types of text apart from caption and scene text shown in Figure of Section 1.2.3. As a result, it is necessary to develop a method for classification of different text types because it is hard to develop single or a unified method for many type text detection and recognition. The main reason is that as text appearance changes, the shape of character changes. This makes a problem for feature extraction to achieve better results.

- It is a fact that due to contrast variations, the image can have different contrast and resolution. This results in disconnections in the character components and loss of significant information. Therefore, there is a need for increasing the resolution of text in the video or images for improving the accuracy of text detection and recognition.
- In sports and marathon images, text detection is difficult because of complex background and limited text information. This combination makes detection and recognition of text more challenging.
- Since text detection is pre-processing step, it detects both caption and scene text well. However, recognition is not as easy as text detection. The main reason is scene text can exhibit any characteristic while caption text has fixed characteristics. Therefore, it is required to classify the caption and scene text before recognition.
- When the shape of character changes along with low contrast, and orientation, it is difficult to apply a binarization method for recognition. Therefore, it is required to implement a robust method without binarization.

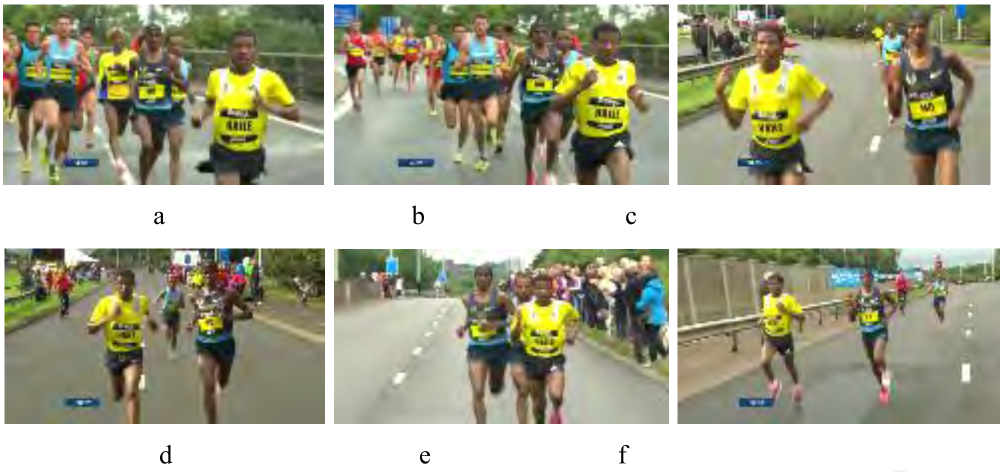


Figure 1.12: Challenges in marathon video text detection & recognition, (a) and (b) contain horizontal text, (c) and (d) contain non-horizontal text, (e) and (f) contain occluded and distorted text.

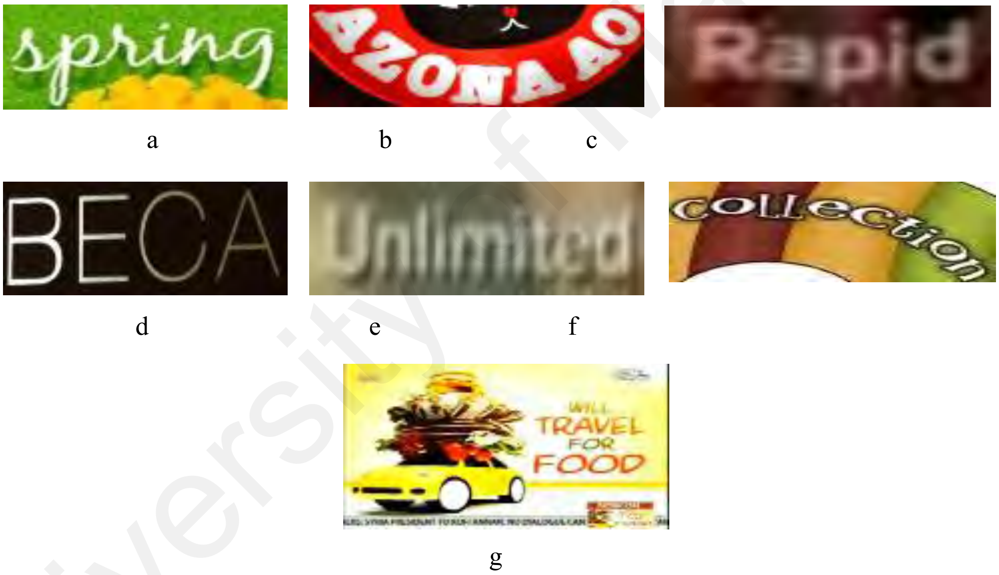


Figure 1.13: Challenges in video text recognition, (a) Cursive text (b) Curvy Text (c) Text with low resolution (d) Text with variant illumination (e) Text with Artifact (f) Curvy text with complex background, (g) Multi-type text (caption and scene).

1.4 Objectives

To address the above challenges mentioned and discussed in the previous section, the following objectives are set to achieve the goal.

i. To overcome the issue of different text appearance of video according to different applications, the combination of rough-fuzzy is explored for classification of video frames of different text.

ii. To increase the low contrast text information, a general enhancement algorithm is introduced using fractional calculus for video/images.

iii. To ease the background complexity, multimodal approach such as face, skin and spatial approach such as texture and context on enhanced video is investigated such that subsequent steps work well.

iv. To reduce the complexity of recognition in video, DCT coefficient is explored for separating scene and caption text.

v. To prevent the loss of shapes from the binarization, Bayesian-based probabilistic approach is introduced for recognition of text. Furthermore, to expand the method for the different scripts, the fusion of SVM and HMM is exploited for scene/video text recognition.

1.5 Contributions

This thesis investigates an efficient text recognition techniques in the video. The contributions of this thesis to video text recognition are listed below and address the problem discussed in Section 1.3. The overview of contribution is shown in Figure 1.14.

Contribution 1: First as one of the major contribution of the thesis, a classification method has been proposed for classifying text frames from different video types. When the text components share the same characteristics, this leads to confusion, uncertainty. To alleviate this issue, the proposed work combines the rough and fuzzy for classification of video frames.

Contribution 2: Exploring fractional calculus for enhancing the low contrast text information distorted by Laplacian operation is new. Despite Laplacian operation enhances the low contrast text information, it introduces noise. To reduce the effect of noise, a fractional calculus-based model is introduced which suppress background and noise such that edges get sharpened.

Contribution 3: For the images like sports and marathon, to reduce the complexity of the background, a multimodal approach which combines face, torso detection is proposed to increase the performance of text detection and recognition. The way the proposed work utilizes the advantage of the face, torso detection is the main contribution.

Contribution 4: Fourth, a classification method has been proposed to classify caption/graphics/superimposed and scene texts in video frames for addressing text recognition which can significantly improve the recognition performance. This classification scheme explores frequency domain features instead of pixel information in spatial domain which is not good in handling noise and distortion. This module explores Discrete Cosine Transform and wavelet coefficients of sub-bands at different levels for detecting text candidates.

Contribution 5: Finally, the fifth contribution as a major purpose of thesis aims to develop recognition method which can handle contrast variation, multi-type of video text, variant font style, orientation, and scripts. To achieve this, the Bayesian classifier is explored for recognition through binarization. To prevent the loss of information by binarization, a method without binarization is proposed based on fixing automatic window detection to extract statistical and texture features in contourlet wavelet domain. The proposed method utilizes spatial and frequency domain to preserve the character's shape.

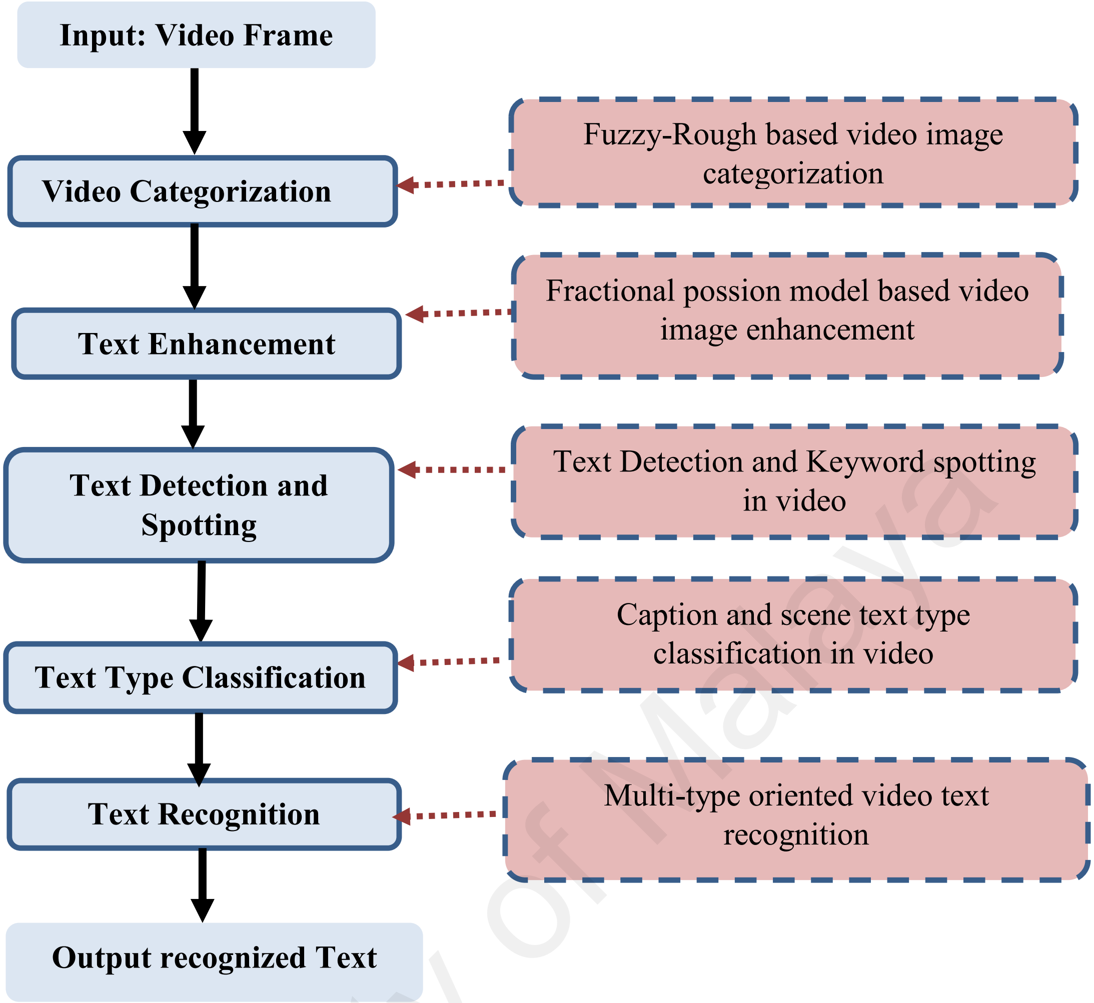

Input: Video Frame
Video Categorization
Text Enhancement
Text Detection and Spotting
Text Type Classification
Text Recognition
Output recognized Text
Fuzzy-Rough based video image categorization
Fractional possion model based video image enhancement
Text Detection and Keyword spotting in video
Caption and scene text type classification in video
Multi-type oriented video text recognition

Figure 1.14: Block diagram of contributions to video text recognition.

1.6 Layout of the thesis

The organization of the chapters in this thesis is as follows:

- In Chapter 1, a preview of the whole thesis has been provided including the motivation of undertaking this research, research background, challenges, objectives and contributions.
- In Chapter 2, the existing recognition methods are reviewed in an elaborate manner in order to understand the necessity of employing video classification, enhancement model, video text type classification, and recognition.

- In Chapter 3, a new novel fuzzy-mass based method has been proposed for classifying text frames from different video types. Fuzzy logic has been introduced for identifying straight and curved components in the edge image. For identified edge components, mass-based features and proximity-based features are extracted locally and globally in a new way by drawing consecutive ellipses over the image. Then Support Vector Machine is employed on extracted features for the final classification.
- In Chapter 4, a new enhancement model has been proposed based on Fractional Poisson for increasing fine details in natural scene images as well as video by suppressing the noises introduced by Laplacian operation. The proposed method considers edges and their neighbour information to derive a mathematical model to increase the low contrast information in the scene image as well as in video.
- Chapter 5 presents a multimodal approach for detecting text in the complex images using face, torso detection and then text detection. The detected words are used for spotting text in the video. This step extracts texture-spatial and context-based feature to capture spatial arrangement of pixels in a word. These features together reduce the false positive and improve the word retrieval relevance accuracy.
- In Chapter 6, a new idea has been proposed of exploring Discrete Cosine Transform for identifying tampering information for the classification of caption and scene texts. Furthermore, a new idea of exploring positive and negative coefficients of wavelet decomposition is also explored for detecting text candidates. The distribution of text and non-text candidates over caption and scene word images are studied in a novel for classification.

- In Chapter 7, a new binarization algorithm has been developed for video image by introducing a Bayesian probabilistic technique. Furthermore, a new idea of fixing window is proposed for the character components of arbitrarily oriented words based on the angular relationship between sub-bands and a fused band. For each window, features are extracted in contourlet wavelet domain to detect characters with the help of an SVM classifier. For each detected character, HMM is proposed for recognizing words and characters of any orientation using the same feature vector.
- Conclusions are given in Chapter 8. First, a summary of the main contribution has been given. Finally, future work is reported.

CHAPTER 2: LITERATURE SURVEY

2.1 Background

In the previous chapter, the importance, application, future challenges and objectives of the thesis are explained. This chapter presents the review of each contribution of the work to understand the state of the arts in video and to show that existing methods have inherent limitations to find an accurate solution of detection and recognition to multi-type and multi-oriented text.

2.2 Video Image Categorization

Categorization of images and videos is a common problem in image processing domain. There are mainly four categorization approaches. These are an image without text, an image with text, video without text and video with text categorization. This section presents a review of the above-mentioned categories in the following sub-sections.

2.2.1 Without Temporal Information

In this approach, single frame or scene image is used for classification instead of temporal information. This approach can be divided into two classes, namely, natural scene image/video frame (without text) categorization, and Image/video text (with text) categorization.

2.2.1.1 Scene/Video Frame Categorization

There are mainly three processing steps in standard image classification pipeline: first, extracting spatial cue like texture using local feature, for example, SIFT (Scale Invariant Feature Transform) (Lowe, 2004), HoG (Histogram of Gradients) (Dalal & Triggs, 2005), or SURF (Speeded-Up Robust Features) (Bay, Tuytelaars, & Van Gool, 2006)); second, encoding of local feature descriptors using the bag-of-words (BoW) (Csurka, Dance, Fan, Willamowski, & Bray, 2004) or Fisher vector (FV) (Perronnin, Sánchez, & Mensink,

2010) representation; and finally, training the encoded features by classifier, such as, Support Vector Machine (SVM), Neural Networks (NN). For example, a combination of SIFT and probabilistic approach has been proposed in Bosch et al. (Bosch et al., 2008) for scene image classification. For learning representation about thousands of objects from millions of images, a CNN-based classifier (Krizhevsky, Sutskever, & Hinton, 2012) have exceeded the accuracy of related literatures on many challenging image datasets. Deng et al. (Deng et al., 2009) has obtained the very promising result using large networks trained on ImageNet. They experimented on many standard image classification datasets and with an SVM. Razavian achieved the same result with an SVM even with no fine-tuning (Razavian, Azizpour, Sullivan, & Carlsson, 2014). Sometimes, the model which is trained on one dataset (Krizhevsky et al., 2012) has been successfully transferred to the other dataset. This type of model is used as a mid-level image representation (Girshick, Donahue, Darrell, & Malik, 2014; Razavian et al., 2014). It shows good results in object and action classification and localization (Oquab, Bottou, Laptev, & Sivic, 2014; Razavian et al., 2014). Similarly, for image classification, deep learning (Lazebnik, Schmid, & Ponce, 2006) has also been explored. For image annotation, a large number of features are extracted from cloud computing. The performance of this system relies on the type and shape of the objects present in learning database. Nowadays, deep learning has been explored by many researchers for video classification because it has been proved from many state-of-the-art studies, that deep learning framework solves the complex image classification problem with higher accuracy (Cloud Vision, 2011; Dunlop, 2010; J. Liu, Chen, Zhu, Liu, & Metaxas, 2016; Nogueira, Penatti, & dos Santos, 2017; Ohn-Bar & Trivedi, 2017; D. Tian, Sun, & Vetro, 2016; Z. Xu, Hu, & Deng, 2016). However, most of the methods focus on particular data types, and deep learning frameworks are designed according to requirements (W. Liu et al., 2017). Besides, optimizing parameter and annotating a large samples are the tricky and cumbersome process, respectively.

Therefore, the complex general video image classification problem cannot be solved by these methods.

From the above discussion, it is noted that the methods ignore text information in the images or video image for understanding content. Most of the methods focus on shapes of the objects in the image and consider character as one of the objects for studying the content of the image. However, the methods do not consider advantage of text information, which helps to extract meaning close to content.

2.2.1.2 Video Text Image Categorization

In general, the video text provides information in the form of scene text or sub-tiles/caption which helps in understanding the semantics of the images easily. Therefore, this section reviews the methods on video or scene with text information.

There are a very few text frame classification methods in the literature. These methods help to increase the accuracy of text detection methods. For instance, Qin et al. (Qin et al., 2016) extracted spatial, structural and statistical features for classifying text type images. However, these methods cannot solve the problem when images have uncertainty and ambiguity. In addition, the classes considered in this work are different for classification. Shivakumara et al. (Shivakumara & Tan, 2010) proposed an edge based method to identify text and non-text block. The edge-based features, like, height and straightness of the edges are exploited in the block level. However, the block losses its text properties due to fixed size blocks. Therefore, the classification accuracy is not satisfactory for text frame compared to the non-text frame. In another work, Shivakumara et al. (Shivakumara, Dutta, Phan, Tan, & Pal, 2011) proposed a text frame classification based on wavelet and moments at the block-level to identify the text candidate. Although the performance of the method is satisfactory on different orientation, scripts, etc., the accuracy is not good for some text frames. The reason is that very strict conditions are imposed on the blocks for identifying the text.

In summary, the above methods are good for the images but not video because the method ignores vital information of video, such as temporal information for understanding content of the images. In addition, current video categorization approaches do not consider the relationships among video semantic; rather they often generate multiple tags to a video sample.

2.2.2 With Temporal Information

In this subsequent sections, categorization methods that utilize temporal frames will be reviewed.

2.2.2.1 Video Categorization

One straightforward approach considers a video as a set of images and relies on techniques designed for image classification discussed in subsection 2.2.1.1. The progress of video classification using temporal information has been accelerated by robust features like SIFT, HOG, and SURF. For example, HoG feature based on 3D spatial-temporal gradients (Klaser, Marszałek, & Schmid, 2008) has been used for action recognition. Similarly, Harris corner detector is introduced into 3D volumes (Laptev, 2005) to detect space-time interest points due to its great success in 2D application. Wang et al. (H. Wang & Schmid, 2013) presented an approach based on densely sampled local patches for estimating dense trajectory features. This feature is tracked over temporal frames in an optical flow field. Later, quantization techniques like Fisher Vector and Bag-of-Words are combined with the extracted features. This fusion achieved prevailing performance on different types of standard databases (Han, Singh, Morariu, & Davis, 2017; Oneata, Verbeek, & Schmid, 2013). Although this approach can model the local motion patterns for a short period of time, it totally ignores the temporal feature of videos.

Graphical models, such as, Hidden Markov Models (HMM), Conditional Random Fields (CRF), Probabilistic graphical model (PGM), etc., have widely been applied to extract spatial structure and long-term temporal pattern (Tang, Fei-Fei, & Koller, 2012; Vail, Veloso, & Lafferty, 2007; Y. Wang & Mori, 2009). For example, Tang et al. has proposed an approach based on HMM (Tang et al., 2012). They explored a variable duration HMM and latent variables are explored over the frames of a video. The advantage of this model is that it performs fast, which is very important for processing a very large number of videos efficiently and quickly. A graphical model is proposed by Deng et al. (Deng et al., 2009) to encode hierarchy structure for improving classification accuracy. Fernando et al. (Fernando et al., 2015) explored the spatial and motion information. They introduced a learning-based temporal pooling method for capturing the global temporal feature. Wang et al. (L. Wang et al., 2016) modelled long-range temporal structure in a temporal segment network. They combined video-level supervision and sparse temporal sampling technique for efficient and effective learning. Recently, LSTM (Long short-term memory) has been explored in sequential modeling for video classification by many researchers due to its success in speech recognition (Graves, Mohamed, & Hinton, 2013) and video captioning (Sutskever, Vinyals, & Le, 2014). It helps to extract temporal dynamics of a video. Srivastava et al. (Srivastava, Mansimov, & Salakhudinov, 2015) proposed an auto-encoder framework to learn video features. This framework is based on LSTMs. A long-term recurrent convolutional networks (LRCNs) has been proposed by Donahue et al. (Donahue et al., 2015). This approach integrates long-range temporal recursion and convolutional layers. Two LSTM models are used. These are spatial and motion features, extracted from CNN. A multiple-layer LSTM has been proposed by Ng et al. (Ng et al., 2015). In this approach, LSTM is extended to five layers and integrated with several pooling architectures. This type of architecture leverages LSTMs to acquire temporal structure and it avoids the shortcomings of the frame based CNN approach. There are many literature works which work on the fusion of multiple features or multiple approaches to accelerate classification score. Mainly, there are two fusion approaches, i.e., early fusion and late fusion. Early fusion is performed at the feature level whereas later is executed at the classification score level (Snoek, Worring, & Smeulders, 2005), (Y. Yang et al., 2013). Generally, early fusion combines features by simple concatenation (H. Wang & Schmid, 2013) or a linear combination of their kernels (Jianguo Zhang, Marszałek, Lazebnik, & Schmid, 2007) before classification. Additionally, Multiple Kernel Learning (MKL) approach can also be used to combine feature kernels to learn the weights automatically. On the other hand, late fusion approach combines prediction scores from multiple classifiers (D. Liu, Lai, Ye, Chen, & Chang, 2013; G. Ye, Liu, Jhuo, & Chang, 2012). Due to the simplicity, both fusion methods work well. However, the features or prediction scores are assumed to be explicitly complementary to one another; thus these approaches fail to capture potential hidden correlations among features. Recently, Deep Boltzmann Machines (DBM), has been applied by Srivastava et al. (Srivastava et al., 2015) and deep auto-encoder has been exploited by Ngiam et al. (Ngiam et al., 2011). These techniques learn the association among different modalities, for example, image and text. In Wu et al. (Z. Wu, Jiang, Wang, Pu, & Xue, 2014) relationship between feature and class is explored by imposing trace norms. From the above-stated literature, it can be said that the context feature or co-occurrence of semantic information, can provide useful a clue. For example, Rabinovich et al. (Rabinovich, Vedaldi, Galleguillos, Wiewiora, & Belongie, 2007) proposed a CRF model incorporating the semantics context information. The internal relationships of video semantics is captured by Wu et al. (Z. Wu et al., 2014) for regularizing the categorization process. To anticipate the context of a category, Chen et al. (X. Chen & Gupta, 2015) exploited confusion matrix during training of CNN's. To use CNN on video data, the most straightforward strategy is stacking frames as inputs. Using

3D convolutions (Ji, Xu, Yang, & Yu, 2013; Karpathy et al., 2014; Tran, Bourdev, Fergus, Torresani, & Paluri, 2014), this CNN-based approach learns spatial-temporal features. However, these works did not perform well compared to existing trajectory features (H. Wang & Schmid, 2013). To effectively model 3D signals, Simonyan et al. (Simonyan & Zisserman, 2014) explored separately two CNN's to capture spatial and motion structure. Following the same idea, Wang et al. (Xiaolong Wang, Farhadi, & Gupta, 2016) introduced a method which learns the transformation between two states. Different types of fusion strategy are integrated by Feichtenhofer et al. (Feichtenhofer, Pinz, & Zisserman, 2016) to combine spatial and temporal information. Generally, stacked optical flow images are not considered; thus the temporal information of videos are avoided to reduce computation complexities. Wu et al. (Z. Wu, Jiang, Wang, Ye, & Xue, 2016) integrated CNN and LSTM to utilize the advantages of both the approaches. Specially, spatial, short-term motion, and audio clues are modelled in different layer of CNN and long-term temporal dynamics are explored using LSTM networks.

In summary, the current methods use temporal structure but ignore text in the image for categorization. As a result, the methods process a large number of temporal frames using complex operations; thus become expensive.

2.2.2.2 Video Text Image Categorization

This section reviews the methods that use text information for categorization.

Dimitrova et al. (Dimitrova, Agnihotri, & Wei, 2000) has classified a TV program based on text/face. The video clip is represented as a series of frame labels using HMM method. The label strings as observation sequences are used to train HMM's. The final model estimates the probabilities or confidence score of the test clip being one of the four categories of TV programs (news, commercials, sitcoms and soaps). Shahraray and D. Gibbon (Shahraray & Gibbon, 1995) explored two information, mainly, key-frames and text information of news video. In this approach, automatic structural and content analysis are not required. Xu et al. (C. Xu, Wang, Wan, Li, & Duan, 2006) have proposed a novel framework for event detection incorporating web-casting text in sports video. The authors have extended their work in (C. Xu et al., 2008). Zhang and Chang (D. Zhang & Chang, 2002) have detected baseball video event using caption text detection and recognition. A integrated method using text and visual features has been proposed by Brezeale and Cook (Brezeale & Cook, 2006) for video classification. This approach extracts closed captions from DVDs and caption are represented as term-feature vectors. Finally, an SVM is used for classification. There are some works which have used multi-modality approach for video text categorization (Evangelopoulos et al., 2009; Q. Huang, Liu, Rosenberg, Gibbon, & Shahraray, 1999). For instance, Huang et al. (Q. Huang et al., 1999) has integrated audio, video, and text for automatic generation summary of news clips. Qi et al. (Qi, Gu, Jiang, Chen, & Zhang, 2000) explored a combination of audio, visual and text in a video. This fused technique extracts content and structure information of video. A new frame classification method has been proposed in (N. Sharma, Shivakumara, Pal, Blumenstein, & Tan, 2015) using linear and non-linear properties of the text component. Manisha and Sharmila (Manisha & Sharmila, 2016) have classified text frame in video sequence before feeding it into detection and recognition method to increase accuracy.

In summary, because of arbitrary text movement, variation in background, low contrast and low resolution, the methods may not work well for the arbitrary orientated text. In addition, there is no proper method reported in the current literature to decide the frames needs to be processed for categorization.

2.3 Video Image Enhancement

In general, video suffers from low resolution. There are many methods proposed for video and image enhancement which can be classified as image enhancement where they focus general images, video enhancement where they explore temporal information for enhancement and video text enhancement where the method focus on text in the image.

2.3.1 Image Enhancement

There are a number of image enhancement approaches reported in the literature. The most popular approach is the Unsharp Masking (UM) technique (Polesel, Ramponi, & Mathews, 2000) for image enhancement. The approach consisted of three basic steps. First, a low-pass filter is applied on target image to make it blurred. After that, the disparity between blurred and observed image is calculated. Finally, a part of the disparity is added to the target image. Although this method is robust to recover the sharpness of the image, it introduces unwanted ringing artifacts in the image. There is another class of enhancement method, which is shock filter (Guichard & Morel, 2003; Rudin, 1987). In this approach, image edges are enhanced by finding the solution of an inverse partial differentiation. A dilation and erosion based process has been proposed by Osher and Rudin (Osher & Rudin, 1990). This approach sharps the images. However, due to the morphological operations, the enhanced image becomes piecewise constant. There are many other types of shock filters designed by (Guichard & Morel, 2003; Schavemaker, Reinders, Gerbrands, & Backer, 2000; Weickert, 2003) to enhance edges. The shock filters can generate a highly sharp enhanced image; however, it sharps all the detected edges irrespective of their original sharpness.

Recently, there have been many enhancement works which are based on super-resolution of a single image. The techniques of super-resolution are classified into three categories, namely, interpolation, reconstruction, and learning technique. In, the interpolation-based approach (Hou & Andrews, 1978; X. Li & Orchard, 2001; Thévenaz, Blu, & Unser, 2000) surrounding known pixels interpolate the value of unknown pixels to increase the resolution. Recently, some authors have made a sophisticated interpolation using the sparse mixing estimation (Mallat & Yu, 2010) and 2-D autoregression

(Xiangjun Zhang & Wu, 2008). However, the interpolation-based approach blurs the high-frequency details if image has jaggy artifacts. The second, reconstruction based approach (Baker & Kanade, 2002; Ben-Ezra, Lin, & Wilburn, 2007; Irani & Peleg, 1993; Lin & Shum, 2004) imposes a reconstruction constraint. This entails the down-sampled and smoothed images, generated from the high-resolution (HR) image. A typical reconstruction based method is Back-projection (Irani & Peleg, 1993). It creates jaggy or ringing artifacts in the vicinity of edges because no regularization parameters are constrained. Therefore, for reducing these artifacts, it is necessary to regulate the reconstruction constraint. Nowadays, machine learning approach has been applied to single image super-resolution (SR) using low-resolution (LR) and high-resolution (HR) image. The co-relation or association between LR and HR patches has been learned by machine learning. In the learning-based approach, high frequency details are "hallucinated" from HR/LR image pairs, present in a training set (Chang, Yeung, & Xiong, 2004; Fattal, 2007; Freeman, Jones, & Pasztor, 2002; Freeman, Pasztor, & Carmichael, 2000; Jianchao, Wright, Huang, & Ma, 2008; Q. Wang, Tang, & Shum, 2005). In (J. Sun, Zheng, Tao, & Shum, 2003), primal sketches such as, corners, edges, and ridges, are hallucinated because these features are more susceptible for recognition to the human eye. The main disadvantage of the learning-based approach is that the number and types of training data determine the performance of the system. The selection of a number of training examples is fuzzy to the generic images. An edge statistics using Edge-Frame Continuity Moduli (EFCM) has been proposed by Fattal (Fattal, 2007). For image up-sampling, the learning of EFCM is done between HR and LR pairs. Various learning algorithms have been explored to learn the relationship between LR to HR. Some of learning algorithms are nearest neighbor approaches (Freeman et al., 2002), manifold learning (Chang et al., 2004), sparse coding (C.-Y. Yang, Huang, & Yang, 2010; J. Yang, Wang, Lin, Cohen, & Huang, 2012; Zeyde, Elad, & Protter, 2010), and convolutional networks (Dong, Loy, He, & Tang, 2014).

However, the learning methods have some disadvantages. First, the selection of types and number of images for training are not clear. Therefore, a training set consisted of wide variations data are often required to learn a vast LR-HR dictionary. Second, for every new type of data, the model needs to be retrained using sophisticated learning algorithms. To avoid problems associated with learning methods, internal patch redundancy is exploited by several approaches (Ebrahimi & Vrscay, 2007; A. Singh & Ahuja, 2014) for SR. The fractal properties of images are extracted from the image (Barnsley, 2014). Ebrahimi and Vrscay (Ebrahimi & Vrscay, 2007) fused concept of example-based approaches and fractal coding (Barnsley, 2014). Non-local means filtering (Buades, Coll, & Morel, 2005) is an example of this approach. The author explored self-similarity technique. Similar to this, Michaeli and Irani (Michaeli & Irani, 2013) applied self-similarity to blur kernel and recover the HR image. Singh et al. (A. Singh, Porikli, & Ahuja, 2014) extended the self-similarity approach to convert the noisy image into SR.

However, through self-similarity matching, these methods are not sufficient to handle multiple planes and recover regular textural patterns. In addition, these methods do not perform well in both video, and natural scenes in one framework as these are not a generic SR algorithm.

In summary, it is noted from the review that strong attention has been given in modeling natural image in the last 2 decades. But it is a fact that the resolution of the video is much lower than scene image due to a variety of capture mediums as seen in the Introduction section. Therefore the method may not work well for video.

2.3.2 Video Enhancement

In this section, the existing works of enhancement from the low-resolution video are reviewed. Various super-resolution methods have been explored on low resolution input videos. For enhancing the resolution, some have introduced in the frequency domain (Tsai, 1984), and some have designed in the spatial domain (Keren, Peleg, & Brada, 1988), and there exist methods which combine both approaches (Demirel & Anbarjafari, 2011). In video enhancement, registration plays an important role. The researcher has investigated many techniques to obtain a good registration; thus subsequently, they improved resolution enhancement. For example, Reddy and Chatterji (Reddy & Chatterji, 1996) has explored an approach using frequency domain for increasing the resolution. A motion estimation technique has been proposed in (Irani & Peleg, 1991). In this approach, translations and rotations are considered for video enhancement. A frequency domain technique has been presented by Cortelazzo and Lucchese (Lucchese & Cortelazzo, 2000) for estimating planar roto-translations. Vandewalle et al. (Vandewalle, Süsstrunk, & Vetterli, 2006) also exploited in a frequency domain. In their method, images were represented by a planar motion.

An important role is played by high frequencies in resolution enhancement in video. Therefore, it is useful if high frequency is separated from low frequency. Wavelet transform is one of the widely used tools especially in video super-resolution techniques (Anbarjafari & Demirel, 2010; Piao & Park, 2007). Mallat (Mallat, 1999) proposed a discrete wavelet transform (DWT) in a video sequence. This one-level DWT operates on a single frame at a time. DWT yields a low-frequency sub-band, and three high-frequency sub-bands. These sub-bands are oriented at 0◦, 45◦, and 90◦. Similar to DWT, there is another approach, named as Stationary video transform (SWT). Unlike DWT, SWT does not down-sample the input; hence the input and output sub-bands have the same size. Anbarjafari et al. (Anbarjafari, Izadpanahi, & Demirel, 2015) proposed a new method using DWT and SWT for enhancing video resolution. Before registration process, they applied an illumination compensation technique and then used SWT and DWT. They applied these two-transform techniques for maintaining the high-frequency part.

From the above review, it is noted that the same enhancement methods cannot be used for video text enhancement because these are applicable to general video but not on video text. In following subsection literature survey of video, text enhancement has been reviewed.

2.3.3 Video Text Image Enhancement

A few number of image/video enhancement approaches has been found in literature. As mentioned in previous sub-sections, interpolation method such as cubic-spline interpolation is most familiar technique used in enhancement. Interpolation method mainly creates two problems. First, smoothing is indiscriminate. It causes blurring in sharp edges of text. Second, this approach is unpredictable. Therefore for maintaining consistency and ensuring distinct boundaries, a robust model is needed. A deblurring method has been proposed on scene text in video by Li and Doermann (Huiping Li & Doermann, 2000) using projection technique. This is especially suited for caption text as these texts have clear translation between temporal frames in video. Similarly, a projective transform motion model in image sequences has been presented by Capel and Zisserman (Capel & Zisserman, 2000) for increasing resolution of text. In both the approaches (Capel & Zisserman, 2000; Huiping Li & Doermann, 2000), the authors improved the images quality. For enhancing text edges, Mancas-Thillou and Mirmehdi (Mancas-Thillou & Mirmehdi, 2005) adapted Teager filter. This filter extracts the high frequencies in a video frame. Any text image property is not considered by most of these prior models. A text-specific prior model has been explored by Donaldson and Myers (Donaldson & Myers, 2005) using local smoothness with step discontinuity. Although piecewise smoothness reduces false speckles in video, but it fails to enhance edges of text. A training-based, Bayesian framework was employed by Dalley et al. (Dalley, Freeman, & Marks, 2004). Park et al. (J. Park, Kwon, & Kim, 2005) proposed an edge-based method for enhancing resolution. Initially, the method detects the edges to sub-pixel accuracy in video. Then the information of compressed edge is fused into the super-resolved image with the help of an MRF formulation. For improving the contrast of a video text frame, Sattar and Tay (Sattar & Tay, 1999) integrated fuzzy edge detectors and a multi-resolution pyramid. A nonlinear optimization was investigated on a grayscale image. This optimization minimizes a Bimodal Smoothness Average rate. The smoothness function performs well due to its robust property. However, edge and texture especially in the corner of text are not preserved by this smoothness function. Here, a differential equation based method that process direction of the smoothness constraint is characterized by the gradient magnitude. In this approach, the main advantage is that random attribute is not considered in the image. Generally, this technique generates a negative results around the corners and edges of character. Existing methods (Ayyalasomayajula & Brun, 2014; C. Chen, Zhang, Bu, Wang, & Chen, 2010; Howe, 2013; Phan, Shivakumara, & Tan, 2009; Caijuan Shi, Ruan, & An, 2014; Shivakumara, Suhil, Guru, & Tan, 2014) have been exploited based on gradient operation with Laplacian mask to enhance text information in video images. Generally, Laplacian helps in identifying abrupt changes from background to foreground and vice versa. It provides high peaks in positive and negative direction in those changes. This peak information is useful for detecting, and recognizing text as this information is the basis for extracting features to detect text and separate foreground (text) from the background in binarization. By integrating stroke-width transform (SWT) and Grab-cut method, Bosamiya et al. (Bosamiya, Agrawal, Roy, & Balasubramanian, 2015) proposed an enhancement method in text video.

In summary, from the above discussion, it can be inferred that there is no consistent enhancement method for reducing the noise effect of Laplacian operation. This shows that there is no generalized model for reducing the effect of Laplacian operation.

2.4 Text Detection and Spotting in Video

In the previous section, a detailed review has been presented on enhancement methods for video. This section presents a review on text detection and word spotting which are required for text recognition in video.

2.4.1 Text Detection

Numerous approaches have already been explored on text detection from natural scene images. This section also reviews the literature on text detection in video. The review also includes the synopsis of improvement made in the field of horizontal, non-horizontal, multi-oriented, multi-types text detection techniques from video.

2.4.1.1 Text Detection in Natural Scene Image

As an example, a method is proposed by Epshtein et al. (Epshtein et al., 2010) on text detection in natural scene images using stroke width transformation. The method detects text following the assumption that the width will remain constant over all character-components and then the eliminations are done based on the extracted text properties. The stroke-width method can detect horizontal and slightly inclined but not multi-oriented and multi-typed ones. Similarly, in (C. Yi & Tian, 2012), the author explored text detection using stroke segmentation, clustering, and string classification in scene images. To filter out background interferences, the approach fuses stroke segmentation based on the fusion of structural and color aspects of the stroke, clustering and text-segment classification based on the Gabor based features. However, it might not perform well in low resolution image due to its dependency on color.

In (Yao, Bai, Liu, Ma, & Tu, 2012), the author extracted geometrical features and stroke-width from connected components for identifying probable text entries of any arbitrary orientations. Finally the actual text was identified based on the candidate linking and chain analysis methods. Having addressed the orientation aspects, this method was not able to find solution for multi-script text, whereas Gomex and Karataza et al. (Gomez & Karatzas, 2013) introduced an approach by investigating the MSER and single linkage clustering for detecting multi-script text. Despite its robustness in detecting multi-script text, the way value chosen in this paper may not be applicable for arbitrary orientations and other languages. Color External Regions and neural network based generalization had been proposed by Sun et al. (L. Sun, Huo, Jia, & Chen, 2014) for text detection. However, there lies a challenge to apply this supervised method which is again limited to horizontal text, for a generic and real-life application like multi-script detection, etc. In (Yin et al., 2014), the author proposed a method exploring MSER to find possible test candidates. Single-link clustering technique was used to link probable text candidates. Then the method eliminates false positives using a classifier. The method works well for scene and Born digital images. However, the applicability is limited to the horizontal direction. A robust text line detection method was proposed by Kang et al. (Kang, Li, & Doermann, 2014) based on the higher order correlation based partitioning of MSER and regularization method. Finally, for eliminating false positives, a texture-based classifier is used. Similar to other MSER based method, this method can only detect straight text line. For two-level hierarchical method has been proposed by Rong et al. (Rong et al., 2014) where connected components are identified through adaptive thresholding, followed by SVM based classification of text segments. This supervised, adaptive threshold based method might not perform well for complex video text images directly. A random forest-based unified framework has been proposed by (Yao, Bai, & Liu, 2014) for segregating text and non-text. This supervised technique performs poorly for multi-oriented text images. The method proposed by Wang et al. (L. Wang et al., 2014) depends on the identification of user's intention specific connected components by adaptive thresholding. In their work, AdaBoost classifier was employed to categorize text candidates based on the geometrical and texture related features. Although the method has less computational overhead, it does not take care complex background related interferences. Besides, the method can only perform horizontal text direction. Wang et al. (Xiaobing Wang, Song, Zhang, & Xin, 2015) introduced a multi-layer segmentation and CRF for text detection. The method developed by Yuan et al. (Yuan, Wei, Liu, Zhang, & Wang, 2015) used the dimensions and distance related features of candidate regions. The method extracts full text lines using seed candidate regions. Finally, for filtering false positives, a sparse classifier is used. The robust hybrid approach introduced in (Z. Zhao, Fang, Lin, & Wu, 2015), depends on the generation of confidence map through learning-based partial difference equations. Specifically the map is generation via adaptive binarization technique along with cluster-based analysis. Simple color symmetry based rules govern the detection of text in scene image. However, adaptive binarization makes it sensitive to background complexity. For multi-orientation scene text detection, Yin et al. 2015 (Yin, Pei, Zhang, & Hao, 2015) proposed adaptive clustering. The possible text candidates are grouped using this technique. A simple combination of various characters cues has been used by Agarwal et al. (A. Agrawal, Mukherjee, Srivastava, & Lall, 2017) for creating bound boxes around text regions. HOG, PHOG and variance of stroke widths precisely refine text regions created by edge enhanced MSER. The Fully Convolution Network-based method (FCN) mentioned by Zhang et al. (Z. Zhang et al., 2016) was trained to predict the saliency maps of text regions. The combination of saliency map and character components formed the basis of their text line hypothesis. Recently neural network model has been explored by Liao et al. (Liao, Shi, Bai, Wang, & Liu, 2017) for text detection. A novel image operator has been proposed by Zheng et al. (Yang Zheng et al., 2017) to detect as well as localize text in the images. The job is accomplished by identifying characters through learning of two classifiers, followed by elimination of wrongly identified ones through recursive search algorithm. The method also deals with pruning of repeating components by combining component trees and recognition results.

An integrated method combining text line entropy-based features with CNN model has been developed to verify line segments (Yixing Zhu & Du, 2018).

In summary, the text contrast is high in scene image; therefore, the character-shape is assumed to be preserved in most of the situations. However, in case of video, this does not hold good always. For video, distorted shapes, any kid of disconnections or loss of pertinent information are expected due to poor illumination or contrast level. Hence text detection algorithms, proposed for the high-contrast static image are not exactly applicable for video-text detection.

2.4.1.2 Text Detection in Video

The review of text detection can be categorized into (1) horizontal (2) non-horizontal (3) multi-type (4) multi-type oriented text detection.

(a) Horizontal Video Text Detection

Cai et al. (Cai, Song, & Lyu, 2002) proposed the color and edge information based text detection method. It performs well for captions but performance deteriorates with font size and orientation. Generally, more false positives occur from the edge and gradient-based methods due to heuristically chosen values, employed for separating non-text and text pixels. A text detection method has been proposed by Wong and Chen (Wong & Chen, 2003) based on analysis of intensity values using gradient and statistical based modeling. This method fails to make the group of text and non-text components. For example, Jung et al. (Jung, Kim, & Jain, 2004) observed that geometric features are almost used by most existing works for segmentation of text in the video frames. Since most of the character components have uniform color in text lines, it plays a key role in the text detection. A sequential Monte Carlo approach has been proposed by (D. Chen & Odobez, 2005) in video for detecting text. Initially, text regions are segmented based on Otsu threshold, them segment wise pixel distribution is used for separating text pixel from background. Liu et al. (C. Liu, Wang, & Dai, 2005) extracted statistical features from the variant directions of edge image. For segregating text and background pixels, hard clustering technique namely K-means is applied. Finally, text pixels are grouped and line segments are extracted using geometrical properties in video frames as well as in images. Shivakumara, Phan, and Tan (Shivakumara, Phan, & Tan, 2009) applied rule-based technique to extract of text related information from video. These rules are obtained from the edge image. In (Y. C. Wei & Lin, 2012), gradient differences has been performed for three images which includes input image as well as two downsized images generated from the input one. Finally, in order to identify true text pixels from the text clusters, the SVM classifier has been applied. A fusion of color, gradient and logGabor filter has been proposed by Zhang et al. (Z. Zhang, Wang, & Lu, 2014) for enhancing video text. Then character segmentation is achieved by analyzing vertical properties before text extraction. However, this approach is not suitable for texts oriented in arbitrary direction. Based on SURF features, Yusufu et al. (Yusufu, Wang, & Fang, 2013) proposed a video text detection. For identifying text candidates, this method explored edge and morphological information. Motion estimation helps in identifying moving text blocks. The text blocks are tracked in two successive frames when the features are found to be identical. Here, caption texts are tracked in the horizontal direction but not scene texts. Zhang and Kasutri et al. (Jing Zhang & Kasturi, 2008) devised a method based on character and link energies. Character component-wise energy has been defined in terms of stroke width distance, as it remains constants over all the components. Finally, line extraction both from images or video frames is carried out by traversing maximum spanning tree. The literature revision regarding edge and gradient-based approaches suggest that although these techniques need less processing speed compared to texture-based ones, but results in more fall positives while dealing with complex backgrounds. MSER, color and stroke width similarity measures etc. have been integrated in the method developed by Gomez and

Karatzas (Gómez & Karatzas, 2014)work to detect and track text segments in real time. Since sensitivity of MSER towards blur information is quite high, it deteriorates the overall performance if blur exists in the video frames. Therefore, there are a few approaches which deal with problems of the scene and caption text detection in video. A Canny Text Detector has been presented by Cho et al. (Cho, Sung, & Jun, 2016). It performs well for video texts, but does not consider arbitrary orientation. A multi-resolution processing algorithm has been introduced by Lalita et al. (Kumari, Dey, & Raheja, 2018). The embedded text of different font size are extracted using the algorithm. Finally, SVM is used to filter/extract text more accurately. Very recently, a fully convolutional network with local context is adapted by Chun et al. (C. Yang, Pei, Wu, & Yin, 2018) to localize text.

In brief, most of the methods discussed above consider high contrast scenes or caption in the horizontal direction but not non-horizontal text. However, real video frames or images contain the text of any orientations.

(b) Non-Horizontal Video Text Detection

There are few methods that address the issue of non-horizontal as follows. For example, Huang, Shivakumara, and Tan (Weihua Huang, Shivakumara, & Tan, 2008) explored a motion vector-based approach for detecting text. Although this method works on only horizontal, vertical, but not for other directions. Zhou et al. (J. Zhou, Xu, Xiao, & Dai, 2007) have proposed a method using multi-stage verification as well as connected component-based analysis for horizontal and vertical text detection in the video. It also suffers from orientation issues. For extracting oriented text, a novel method is proposed by Crandall et al.(Crandall, Antani, & Kasturi, 2003). However, similar to the above methods, it only considered graphic texts oriented particularly in 0, 15, 20 degrees and so on. Anthimopoulos and Gatos (Anthimopoulos, Gatos, & Pratikakis, 2013) proposed a method using LBP and a Random forest classifier.

In contrast to component-based approaches, this method is robust enough to handle complex background, but it is computationally heavy.

The above methods are developed for specific text type such as caption text or scene text, born-digital text but not the image with multi-type text.

(c) Multi-Type Video Text Detection

Usually, there are multi-type of text such as scene, caption, and born-digital text, etc. in video and natural scene images. Therefore, this section reviews the methods that address the issue of detecting multi-type text in the videos.

A method exploiting the advantages of both gradient and color space has been proposed by Lienhart and Wernicke (Lienhart & Wernicke, 2002). The input video frame is used to generate the various directions of edge image. Neural network is employed to discriminate non-text and text region, preceded by generation of edge maps from input image. By observing the edge distribution around text, Fourier based statistical feature in RGB space has been explored by Shivakumara et al. (Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010). The method works well for unconstrained background, different scripts, various fonts etc. To find the corner points from video frame, Zhao et al.(X. Zhao et al., 2011) explored Harris corner detector. The shape features around the detected corner points are extracted. Finally, these shape features are employed to formulate the rules to identify caption. An unsupervised clustering scheme has been introduced by Mosleh et al. (Mosleh, Bouguila, & Hamza, 2013). The proposed approach is based on a connected component analysis. The properties of bandlet transform is employed in representing local image geometry. In order to identify both caption and text embedded in video, Wu, Shivakumara, Lu, and Tan (L. Wu, Shivakumara, Lu, & Tan, 2014) devised an approach by exploiting temporal information and Delaunay Triangulation. Identifying moving text is still an issue in this type of approaches. Furthermore, the performance of the method is not satisfactory for multi fonts or multi-sizes in the video. Tracking by detection, contextual learning (spatiotemporal), linear prediction to anticipate probable text location are nicely incorporated in the multi-strategy based text tracking method as mentioned in (Zuo, Tian, Pei, & Yin, 2015). With dynamic programming, Tian et al. (S. Tian, Pei, Zuo, & Yin, 2016) construct a robust and precise text detection system. The dynamic program locates character candidates extensively (learning locally) and explores text regions globally (learning globally). Gaussian-weighted L1 is investigated by Khare et al. (Khare, Shivakumara, Raveendran, & Blumenstein, 2016) in the deblurring model to restore the sharpness of the edges in the vide frames/images. Recently, a approach incorporating Cloud of Line Distribution (COLD) and Random Forest Classifier has been explored by Wang et al. (W. Wang, Wu, Shivakumara, & Lu, 2018). In their work, after extracting the unique shapes of text components, the shape of the texts are studied in a novel way to find the relationship among dominant points over contours of related text segments. It is often referred as COLD in polar domain. An edge is considered as a text candidate if Sobel and Canny share COLD property for that component. Feature extraction occurred in two stages comprising extraction of statistical and angle oriented features and studying the COLD distribution at component level. Next, random forest classifier has been trained on these features to remove false text entries. The cluster is formed to group text lines using distance-based similarity measures between edge components. Finally statistical and angle oriented features at word level are computed to reduce fall positives.

(d) Multi-Oriented Video Text Detection

Shivakumara et al. (Shivakumara et al., 2012) introduced a text detection approach using enhancement in video. The low contrast text pixels are enhanced using Laplacian and Sobel operations, followed by classification of enhanced text pixel using Bayes classifier. The method results in formation of probable non-text and text matrices. Finally, using the nearest neighbor information, a boundary growing approach is employed to span the entire. In (N. Sharma, Shivakumara, Pal, Blumenstein, & Tan, 2012), the author proposed a new text detection method selecting dominant text pixel, and using region growing. Gradient information from Sobel map are exploited to find dominant text pixels. This process avoids under-segmentation of text representatives. Then candidate text grows its perimeter along the text direction. Word patches are formed by grouping the neighboring text components in the Sobel edge. This continuing process extracts variety of oriented text line present in video. A gradient vector flow (GVF) is exploited for the first time by Shivakumara (Shivakumara, Phan, Lu, & Tan, 2013) for detecting text in video. Using Sobel edge map, the method selects dominant text pixels and text candidates. The selection of dominant text pixel removes non-text information in video frames. A wavelet moments-based approach is proposed by Shivakumara et al. (Shivakumara, Dutta, Tan, & Pal, 2014) for detecting text in video frames. Initially, for classifying text frame, they employed a novel symmetry based idea, measured in terms of wavelet and median moments. Then it uses angle projection boundary growing for detecting arbitrarily oriented text. A new multi-modal approach has been presented by Shivakumara (Shivakumara, Raghavendra, et al., 2017) for detecting and recognizing text/bib number in Marathon video. The author employed an upper body detection method along text detection approach to increase the text detection performance. Recently, Yang (X. H. Yang, 2017) introduced network flow based tracking method for detecting text. In their work, text detection is achieved through FCN. Then motion-based method are explored to track detected text in adjacent frames. Next, the method employs a cost-flow network to make association among tracked text. The networks used a min-cost flow algorithm to estimate the text trajectory, followed by a post-processing to improve the performance.

In summary, the detailed literature review suggest that none of the methods deals with multi-type images. To handle the multi-type image, a robust and generic feature extraction is very much required. In addition, when the image contains the multi-type text of multi-oriented, the performance of the method degrades.

(e) Multi-Type-Oriented Video Text Detection

For multi-type video text detection, a technique based on the Laplacian has been introduced by Shivakumara et al. (Shivakumara, Phan, & Tan, 2011) in the frequency domain. First, the Fourier-Laplacian is employed on the input image for filtering purpose. The candidate text region is identified by K-means clustering algorithm based on maximum difference. Text strings are segregated from each other based on the skeleton of each connected component. Finally, for false positive elimination two properties namely, straightness of text string and edge density are used. Gradient-based Delaunay triangulation has been proposed by Wu (L. Wu, Shivakumara, Lu, & Tan, 2015) for detecting and tracking texts placed in arbitrary orientation. For text candidate selection, the technique exploits gradient directional as well as spatiotemporal information at the component level. To evaluate the spatial properties and relationship among corners, Delaunay triangulation is proposed. A novel idea has been proposed by Liang (Liang, Shivakumara, Lu, & Tan, 2015a) based on combination of Laplacian and wavelet. In order to preserve the finer details of text, they have explored MSER and stroke width distances. They introduced mutual neighbor based clustering to group text candidates based on geometrical properties. Finally, the symmetry driven process is iteratively employed to grow the text boundary for extracting oriented text lines. The wavelet and Laplacian combination influenced Liang et al. (Liang, Shivakumara, Lu, & Tan, 2015b) to design a novel algorithm for character segmentation from text line oriented in arbitrary direction. Khare et al. (Khare, Shivakumara, & Raveendran, 2015) proposed a novel feature using Histogram of Oriented Moments (HOM). HOM computes second-order geometric moments, such as values of pixel and spatial information to boost the strength of moments. Edge density and dense corners of component are used to remove false text candidates. In the same approach, the author further modified optical flow for identifying moving text. This method works for detecting static text as well as moving text in video. In another work, Khare et al. (Khare, Shivakumara, Paramesran, & Blumenstein, 2017) introduced a novel text detection approach using gradient and moments directions for multi-oriented and multi-lingual text. The author extracted common stroke width from Sobel and Canny images for finding automatic windows and identifying candidate text pixels. The temporal information is explored by introducing an iterative procedure for candidate text identification. This approach also estimates the starting and ending limit of frames, need to be processed. Moreover, the author used the gradient inward and outward directions of pixels of the edge for eliminating false text candidates. In case of low resolution images such as captured using mobile phones, Shivakumara (Shivakumara, Wu, et al., 2017) proposed another approach based on fractal expansion to detect text for the first time. Self-similarity property of fractals is investigated in a very innovative way to address variant font sizes, multi-script text lines, different types of background contrasts, and arbitrarily-oriented texts. Optical flow is estimated using temporal frames to improve text detection results. Wu (Y. Wu, Wang, Palaiahnakote, & Lu, 2017) explored symmetry and appearance property of the text First, for detecting text candidates in images, the proposed method used Extremal Regions (ER). From each text region, Multi-domain Strokes Symmetry Histogram (MSSH) feature is extracted. MSSH defines the inherent symmetry property of text. To represent text candidate, deep learning network and Auto-Encoder network have been used. Finally, text lines are constructed based on the classification results. Mittal (Mittal, Roy, Singh, & Raman, 2017) has explored histogram of Oriented Moment feature. In this work, SVM and RNN classifier are used for classifying arbitrarily-oriented text/non-text. Yang et al. (C. Yang, Yin, Pei, et al., 2017) followed a dynamic programming approach to solve the problems related to multi-orientation video text detection by using multiple frames. Initially, multi-information based approach is developed to extensively search for probable candidates (as characters) and extraction of test regions. Second, to refine the detection, an optimal tracking trajectory is learned and linked globally for each successive frames using dynamic programming. Fourier-Laplacian filtering in the frequency domain has been proposed by Sain et al. (Sain, Bhunia, Roy, & Pal, 2018). First, in the input image, they employ Fourier–Laplacian transform and Laplacian–Gaussian filtering. After that, K-means clustering is used to find the possible candidate text region using maximum difference map. Next, a skeletonized process has been applied on candidate regions to separate text from false candidate. Finally, a verification of text/non-text is done using HMM.

In summary, it is found from the state of art review on text detection in video, that there are methods which tackle the issues of multi-type oriented images, but these methods do not work on the complex video image.

Therefore, an alternative way to increase the speed of text frame retrieval process based on word spotting which does not require character segmentation and recognition for retrieving. In the following subsection, a detail literature work of keyword spotting in the video is reviewed.

2.4.2 Keyword Spotting

After detecting text in the video, keyword spotting method is used for indexing and retrieving the video from the archive without recognizing through OCR. It is noted from the literature on keyword spotting that keyword spotting is popular work in the field of document analysis where lots of methods for keyword spotting from a scanned document, degraded images, and handwritten document images can be seen. This section reviews the methods on keyword spotting in document images, natural scene images and video images.

2.4.2.1 Keyword Spotting in Document Images

Keyword spotting is a familiar topic in document analysis for document indexing and retrieval. The methods in document image extract feature based on zonalization of words and analyzing connected component to retrieve documents. Lu and Tan (Y. Lu & Tan, 2002) proposed a keyword spotting method from digital document images using the characteristics of connected components and shape codes. This method works well for plain background and high contrast images. Recently, a two-stage method for word spotting in a graphical document is proposed (Tarafdar, Pal, Roy, Ragot, & Ramel, 2013), where rotation invariant features and a SVM classifier are used. The missing characters are restored with the help of SIFT features. For keyword spotting in Bangla handwritten documents, Zhang et al. (Xi Zhang, Pal, & Tan, 2014) proposed a segmentation-free approach using SIFT features. Similarly, a segmentation-free technique has been proposed by Almazan et al. (Almazán, Gordo, Fornés, & Valveny, 2014a). This approach proposes HOG descriptors for representing document images and then uses sliding window procedure for identifying regions near to query words. It uses exemplar SVM for verifying the regions given by HOG descriptor. The scope of the approach is limited to a few variations in handwriting document images. A large-scale graph indexing approach has been proposed by Riba et al. (Riba, Lladós, Fornés, & Dutta, 2015) using binary embedding of node contexts. Here, they exploited a sub-graphs matching to reduce the computational complexity of conventional graph matching. To achieve their objective, they proposed context information for graph matching. However, the proposed method is still considered to be expensive as database size increases due to graph construction and matching. For keyword spotting, Puigcerver et al. (Puigcerver, Toselli, & Vidal, 2015) proposed a probabilistic interpretation method in handwritten document images. This method explores HMM-filter. The HMM-filter involves the probabilistic formulation of keyword spotting. Since these approaches are developed for documents, the same approach may not be used for spotting in videos, natural and license plate images. Wei et al. (Wei, Gao, & Su, 2015) (H. Wei, Gao, & Su, 2015) proposed a multiple instances approach on historical Mongolian document images. The method finds a ranking for query words, and then the rank list is revised by pseudo-relevance feedback to generate an actual ranking of the words. This approach improves keyword spotting. However, the scope of the approach is limited to Mongolian documents. Howe (Howe, 2013) proposed a word spotting method using an Ink ball model. This method combines character recognition modes as string-based queries. Ink ball character model offers an explanatory tool for understanding handwritten marking. The performance of the approach lies in the localization of characters; however, it is not clear how the proposed approach integrates different models. Toselli et al. (Toselli, Vidal, Romero, & Frinken, 2016) proposed a keyword spotting approach using HMM word graph in handwritten document images. This method proposes a holistic and segmentation-free algorithm, which does not require character and word segmentation. Instead, it explores language model to predict posterior probability. However, the use of language model restricts the ability to apply it to different datasets. Mondal et al. (Mondal, Ragot, Ramel, & Pal, 2016) proposed a flexible sequence matching technique in the degraded document for spotting words. This flexible sequence matching combines the advantages of other sequential matching algorithms to make it robust to noises and outliers created by degradation and distortion. However, the performance of the approach relies on text line segmentation and word segmentation. Retsinas et al.(Retsinas, Louloudis, Stamatopoulos, & Gatos, 2016) proposed a keyword spotting using projections of oriented gradients. This technique works on handwritten documents. This approach introduces local and global descriptors based on Radon and Fourier transforms for extracting features. To define the descriptors, the approach explores gradient information in different directions. Thus the approach works well for high contrast images but cannot give a satisfactory result on low contrast. To spot the correct keyword in the handwritten document, a Bayesian background model has been proposed by Kumar and Govindaraju (G. Kumar & Govindaraju, 2017), This approach used an efficient rejection criterion to refine the searching. The inferences of parameters are made based on variational methods and Markov chain Monte Carlo method. However, the method requires more labeled samples and character level scores for achieving good results.

Recently, use of deep learning is getting popular for keyword spotting in handwritten documents as deep learning helps in solving a complex problem. For example, Sudholt and Fink (Sudholt & Fink, 2016) proposed PHOCNet based deep learning for word spotting in handwritten documents. It is shown that how deep learning can withstand the variations of handwriting. Zhang et al. (S. Zhang, Liu, & Qin, 2016) proposed a word spotting approach using a deep neural network where it is shown that how deep neural network can help to spot the words in speech processing. For handwritten keyword spotting, Wicht et al. (Wicht, Fischer, & Hennebert, 2016) proposed deep learning features based on HMM and Dynamic Time Warping (DTW). Sharma (A. Sharma, 2015) proposed a CNN for spotting and recognizing word where the method is tested on different scripts printed documents. It is true that the performance of the deep learning-based approach has been improved significantly compared to existing works but the following concerns are big question mark for the real-time applications and the data affected by multiple adverse factors (A. Sharma, 2015), e.g. (i) determination of a number of annotated samples in training to capture the variety of features at various levels and (ii) optimal design of CNN architectures and their training strategies in term of layer selection are most challenging and tricky task.

Nonetheless, the scope of the aforementioned methods is restricted to scanned document and handwritten document images, respectively, but not for scene image or video. Therefore, it is inferred that performance of existing works is satisfactory for plain background and high resolution document where character shapes are preserved.

2.4.2.2 Keyword Spotting in Natural Scene Images

There are a few approaches which address the issues of keyword spotting in natural scene image. In (Mori & Malik, 2003) the authors introduced a novel technique of breaking visual CAPTCHAs. Almazan et al. (Almazán, Gordo, Fornés, & Valveny, 2014b) explored a word spotting and approach using embedded attributes. This approach considers supervised learning with HMM model for word spotting in the images. The performance of the approach depends heavily on learning and predefined labeled data. As a result, the approach may not be used for the images which often have contrast variations as in video images. In addition, the approach assumes words are available. Wang and Belongie (K. Wang & Belongie, 2010) proposed word spotting in the wild. This approach extract features using local descriptors and then it uses lexicons for spotting the words in the images. A large number of lexicons is needed for isolating text from non-text information in this approach. Since the approach uses local descriptors, the approach may not work well for low contrast images like video images as these descriptors give good features when the image has high contrast. Zamberletti et al. (Zamberletti, Gallo, & Noce,

2015) used augmented text character and CNN for text spotting in scene images. Here, the author explored MSER with augmented reality concept and convolutional networks. The approach uses augmented concept for patch construction, and the patches are used for feature extraction, and classification of text. The proposed approach is tested on both license plate and natural images. This approach requires a sufficient number of training instance in order to obtain satisfactory results. Moreover, the approach has not been tested on video frames. A deep CNN has been proposed by Jaderberg et al. (Jaderberg, Simonyan, Vedaldi, & Zisserman, 2016). A random forest approach is employed to eliminate false-positive candidate word.

From the above review, it can be seen that there are few methods for spotting words in the natural scene images, but most of the approaches assume words are available for spotting. Besides, the performance of the approaches depends on a large and wide range of training data. In addition, none of the approaches are tested on low contrast images like video images containing multi-oriented texts and multi-script images.

2.4.2.3 Keyword Spotting in Video Images

There are a very few existing works reported in the literature on keyword spotting in video images. In (K. Wang & Belongie, 2010), the author introduced an automatic facial expression spotting method in the video. It spots the facial expression among numerous facial expressions without training the model; thus it saves the time of high-level facial analysis. Inspired by this, there are a few spotting works in the video published recently. Roy et al.(P. P. Roy, Bhunia, Bhattacharyya, & Pal, 2017) presents a novel word spotting framework using dynamic shape coding for text retrieval in video frames. They use two-step word-spotting approach consisted of HMM for extracting contextual information from the neighbor to enrich the shape feature. In addition, very recently, a date spotting technique has been proposed by Roy et al. (P. P. Roy, Bhunia, & Pal, 2018) using HMM.

This method spot the date online. Initially, date information is detected using HMM in a given text. Without segmenting characters or words, the author examined different date models from a line. They converted RGB text line to gray image using an efficient algorithm to enhance the text information. After that, binary and gray images are used to extract the Pyramid Histogram of Oriented Gradient (PHOG) feature. The features from the two image are fused with MLP based Tandem technique. Lastly, during word spotting, they exploited a shape coding scheme to boost the performance. This scheme integrates the related shape or silhouette of the same class. These approaches are applicable to the horizontal word but not multi oriented word.

In summary, there are a very few spotting at work in video, and they all assume the text line are available and are in the horizontal direction. Besides, similar to spotting in the natural scene, the performance of above learning-based approaches rely on training samples size. Moreover, none of the algorithm aims to handle video images containing multi-oriented texts.

2.5 Video Text Type Classification

Since video contains multi-type text such as caption, scene, and born-digital text, to improve the recognition performance, there is a need for classifying text type such that an appropriate method can be used enhancing the performance. There are methods that classify printed text from handwritten text to improve the performance of recognition. The following sub-sections discuss a review of text type classification methods in document and video as well.

2.5.1 Printed and Handwritten Text Classification

For separating the handwritten text from a machine-printed text, Pal and Chaudhuri (U Pal & Chaudhuri, 1999; U. Pal & B. B. Chaudhuri, 2001) presented a novel method in two renowned scripts, namely Bangla and Devanagari. They explored feature extraction using statistical and structural concept. Projection profile analysis separate horizontal or vertical text lines in case of line segmentation. Finally, a tree classifier is employed using structural features for achieving the classification. In Kavallieratou and Stamatatos (Kavallieratou & Stamatatos, 2004), the lower and upper profiles of a character are examined using horizontal projection. Then, discriminant analysis is employed to separate handwritten or machine-printed text. A sliding window-based technique has been explored by Santos et al. (Dos Santos, Dubuisson, & Bortolozzi, 2002). They extracted content and shape-related features on each sliding window for identifying the handwritten part in bank checks.

For word-level classification, Guo and Ma (Guo & Ma, 2001) proposed a connected component (CC) based approach. In this approach, the image document is segmented into word-blocks using CC analysis. After that, those word-blocks are consolidated following a set of rules. A projection profile is calculated for each word-block. Then, the profile feature is fed into HMM for classifying machine-printed text or handwritten. Da Silva and Conci (da Silva, Conci, & Sanchez, 2009) built a system for classifying various types of forms, such as questionnaires, subscription forms, or preprinted memorandums. In beginning, the connected component is analyzed to segment form into the word-blocks. From each word-blocks, eleven features are extracted. Using pre-determined thresholds obtained from off-line database, each word block is labelled. A Gabor filter-based approach has been used on word blocks by Farooq et al. (Farooq, Sridharan, & Govindaraju, 2006). Then a neural network is employed for identification of handwritten Arabic text. The full document image is modeled using Markov Random Field. Peng et al. (Peng, Setlur, Govindaraju, & Sitaram, 2013) separated the documents into three classes. These are handwritten text, machine printed text, and overlapped text. A printed/handwriting text classification method has been proposed by Zheng et al. (Yefeng Zheng, Li, & Doermann, 2004) in noisy document. Initially, the CC is analyzed in a document. After that, based on spatial proximity, word blocks are formed by merging the connected components. They extracted a number of features like Bi-level co-occurrence, Gabor filter, and crossing count histogram from those word-blocks. Fisher classifier is considered for the classification of handwritten, machine-printed or noisy document.

The main disadvantage of the previous methods is that the application scope is very restricted to a single context. Most of the existing approaches depend on word level segmentation. Failure in segmentation stage affects the word level classification. Thus, the applicability of application is specifically restricted to document. In these type of documents such as forms or bank checks, the layout is anticipated. Moreover, the segmentation methods that perform the at the component level, fail to handle the noisy content due to the expansion of component size than actual size.

In summary, the performance of most of the methods is restricted to a document where the background complexity of layout is less compared to video text. Since the resolution and background of a video frame are low and complex, respectively, the reviewed methods may not perform well.

2.5.2 Multi-Oriented Text Type Classification in Video

There are a very few works that address text type classification in the video. A frame-based classification has been performed by Shivakumara et al. (Shivakumara, Huang, Phan, & Tan, 2010) . This classification separates high contrast and low contrast video frame. Two dynamic thresholds are experimentally chosen to separate the frames before applying the text detection on the unclassified frame. They showed that this low/high contrast classification improves the text detection result compared to without classification. Inspired by the work, Raghunandan et al. (Raghunandan, Shivakumara, Kumar, Pal, & Lu, 2016) proposed a method to alleviate this problem of text type classification in the natural scene, born-digital image, video, and mobile image.

Shivakumara et al. (Shivakumara, Kumar, Guru, & Tan, 2014) proposed a fusion technique using Canny, Sobel edge and ring radius transform for segregating scene and caption texts. Temporal information is not exploited by this fusion method. Later, Xu et al. (J. Xu, Shivakumara, Lu, Phan, & Tan, 2014) introduced temporal information to separate graphics and scene texts. However, the method requires full-text line for achieving better results. Recently, gradient direction is introduced by Xu et al. (J. Xu, Shivakumara, Lu, Tan, & Uchida, 2016) for the same purpose. The medial axis of character component, and the gradient direction of edge pixels are investigated in a novel way. Ring radius transform is used to find medial axis of character. Bhardwaj et al.(Bhardwaj & Pankajakshan, 2016) explored tampered features for separating scene text from caption text.

From the above literature review, it can be observed that there is a very few text classification work in the video. In addition, developed are not adequate to address the issues of classification of the multi-type text of different orientations.

2.6 Video Text Recognition

Video text recognition methods use either publicly available OCR through binarization or own classifier with/without binarization. The methods which fall into the first category are based on thresholding technique and the methods which extract scale, rotation invariant features such as, HoG, MSER, SIFT, followed by their own classifier, for example, SVM, NN, are in the second category. The following sub-sections present the past works related to text recognition through OCR and without OCR and their respective categories.

2.6.1 Recognition through Binarization

The standard recognition methods which work through binarization generally use threshold for segregation of foreground and background information. The determination of the threshold is tricky and depends on the characteristics of the text. The methods in (Howe, 2013; Milyaev, Barinova, Novikova, Kohli, & Lempitsky, 2013; Moghaddam & Cheriet, 2010; S. Roy et al., 2012; Su, Lu, & Tan, 2013) recognize text through binarization require complete shapes of characters to achieve better recognition rates. This subsections review the recognition algorithms which work through OCR in the document, natural scene, and videos.

2.6.1.1 Document Text Binarization

There are standard methods which are known as baseline methods for binarizing plain documents image. For documents image binarization, generally global thresholding and local thresholding method have been used. These threshold methods convert the image from gray to binary. In global thresholding method, only one threshold is selected. Otsu method (Otsu, 1979) is one of the well-known global thresholding method. In this type of technique, the basic principle of choosing the threshold is to divide the image gray pixel into two groups such that it maximizes the between-class variance. Otsu thresholding works well when foreground of the image has different intensity values compared to the background. Afterwards, in order to overcome the problem of global thresholding method, local thresholding method has been exploited. In local threshold method, the threshold is calculated based on per pixel and its neighbor information. If the input pixel gray color is higher than the neighbor pixels, then it is considered as white, otherwise, black. Many papers have been proposed to calculate the local threshold-based on some statistics measure. For each pixel, Bernsen (Bernsen, 1986) has computed a local threshold based on minimum, and maximum gray value in neighbor region whereas Niblack ‘s (Niblack, 1985) threshold is calculated in each window using mean and standard deviation. Sometimes the threshold calculation depends on histogram analysis. For example, Kopf et al. (Kopf, Haenselmann, & Effelsberg, 2005) has first estimated dominant text color based on histogram calculated on YUV color space and compared it with neighbor region’s dominant bin. The drawback of this approach is that it cannot differentiate the noisy area and text area. Thus, for the non-text region, the threshold is reckoned. Souvola et al. (Sauvola, Seppanen, Haapakoski, & Pietikainen, 1997) has overcome this drawback by considering that gray values of text pixel and background pixel are near 0 and 255, respectively. This method performs better in case of document image than the previous two methods. However, this type of assumption does not work in case of video image where more variation can be expected in color for the text and background, and this assumption could be reversed for some cases as well. A comparative detail study of these threshold-based methods can be found in (J. He, Do, Downton, & Kim, 2005). The size of a window is a key factor for the above-mentioned algorithms. The character size heavily affects window size. The estimation of threshold becomes error-prone if the size of window is small, than the thickness of text edge. As, when the window is fixed at an inward pixel of character, it yields a small standard deviation but large mean. The reason for this abnormality is that majority of the pixels are located in selected window and these pixels belong to the foreground. Therefore, the local threshold becomes a large value for those ambiguous pixels, and consequently, the central and black portion of thick regions are considered as background. On the other hand, the shape of the character becomes imprecise if the size of window is large. Chen & Wu (Y.-L. Chen & Wu, 2009) has introduced a new adaptive algorithm to generate a binary image using multi-plane segmentation approach. These algorithms are particularly restricted to some small skewed text lines in scanned document images. Since they do not deal with oriented words which are more common in the video image, thus the above-mentioned shortcomings generate inconsistent results for the video. Sauvola's method is modified by Guillaume Lazzara et al. (Lazzara & Géraud, 2014) using a multiscale scheme. Recently, a pixel-based binarization evaluation methodology has been proposed by N-tirogiannis, K. et al. (Ntirogiannis, Gatos, & Pratikakis, 2013) for document images. However, this method is unable to recover missing punctuation marks. For these issues, Biswas B. et al. (Biswas, Bhattacharya, & Chaudhuri, 2014) introduced a global-to-local method. For non-uniformly illuminated and severely degraded documents, Brij M. et al. (B. M. Singh, Sharma, Ghosh, & Mittal, 2014) explored a method which corrects touching and broken characters. Howe (Howe, 2011) has proposed a Laplacian method based on pixel intensity for document binarization. Later, the same author (Howe, 2013) improved the Laplacian algorithm with a selection of parameters automatically. As the initial target of these approaches is document or handwritten image, they fail to handle text variations in scene images, for example, variant changes in stroke width, the same color in foreground-background and outlier edges present in the single image. Su et al. (Su et al., 2013) explored a binarizing algorithm on poor quality document images. A multi-scale framework has been proposed by Moghaddam and Cheriet (Moghaddam & Cheriet, 2010) using adaptive binarization. This work also concentrates on degraded document images. Here, Otsu thresholding is adapted in a novel way depending on the complexity of image. However, this thresholding constraint might not work well for scene and video. Using the concept of retinex theory, Wagdy et al. (Wagdy, Faye, & Rohaya, 2015), proposed a binarization technique. The approach integrates the advantages of both thresholding techniques (local and global) to enhance the quality document. A novel binarization method has been explored by Jia et al. (Jia, Shi, He, Wang, & Xiao, 2016) using stroke symmetry. They applied the structural symmetry pixel to calculate the threshold in neighbor area to subdue the non-text pixels. Very recently, a de-noising method has been investigated by Chen and wang (Y. Chen & Wang, 2017) using the non-local means. Then, based on the histogram, adaptive thresholding is applied. For each pixel, the method uses Rosenfeld's technique to select the threshold. Very recently, Jia et al. (Jia, Shi, He, Wang, & Xiao, 2018) estimated the structural symmetric pixels based on the threshold in neighbor region. After that, a pixel corresponds to the background or not is determined using multiple thresholds-based voting result.

In a global thresholding technique, a major shortcoming is that they don't consider the spatial relationships among the pixels, which create difficulties in the presence of noise. Generally, degradations are purely local in the image, so, global methods face many challenges. Overall, the threshold is selected by considering the center part of text and boundary of the text. However, it does not yields good binarize result when text boundary and text are almost same which is common in scene and video image.

2.6.1.2 Natural Scene Text Binarization

To overcome the local and global thresholding issues, there is an alternative way for enhancing binarization algorithm by preserving the characters shapes irrespective of document and scene image. Generally, hybrid or adaptive is based on a combination of local and global thresholding has been applied on for natural scene text for binarization. The main disadvantage of global and local thresholding is to decide the suitable threshold for full image or each possible sub-images, respectively. Global and local both clue are utilized in Hybrid methods to decide the pixel label. In this technique, the main emphasis has been given on establishing a set of criterion and determining the value of parameters by avoiding one fix threshold or simple statistical measure calculated from neighbor pixels. Following the same idea, Gllavata et al. (Gllavata, Ewerth, Stefi, & Freisleben, 2004) investigated a clustering scheme where each region has to be decided text or non-text region based on estimated parameters from the sub-regions. It first enhances the image using cubic interpolation. Afterward, a color quantizer method, which is used to reduce most dominant color, determines the color of text and background. Then they calculate two color histograms: one is from the center part of the text, and another one exactly below and above of the text. Next, the disparity has been estimated from these histograms to choose the minimum value considered as background, and maximum value considered as text. Finally, K-means clustering has been applied to generate binary text image. Similarly, Chou et al.(Chou, Huang, Lin, & Chang, 2005) proposed a binarization method using decision tree approach. In some works, different color space has been used to perform clustering instead of quantizer method. For example, (Yokobayashi & Wakahara, 2005) uses CMY (Cyan, Magenta, Yellow) for binarization. Later the same author (Yokobayashi & Wakahara, 2006) has modified color clustering using global affine transformation. This transformation is a one type of correlation approach that maximizes the class segregation. There are some papers where clustering based binarization has been introduced. In the paper proposed by Garcia & Apostolidis (Garcia & Apostolidis, 2000), the cluster number has been fixed at 4 to reduce computation complexity. However, these methods only work on high-resolution scene image, not video image. A metric-based clustering has been proposed by Mancas-Thillou & Gosselin (Mancas-Thillou & Gosselin, 2007) using a cosine-based similarity. This similarity is measured by the Euclidean distance. Spatial information is extracted using Log– Gabor filtering. This filtering is used for separating the characters components. But this Log-Gabor filter is specially designed for natural scene image, not video image. Huang et al. (R. Huang, Shivakumara, Yaokai, & Uchida, 2013) proposed scene character recognition with the co-operative multiple-hypothesis framework. This method obtains responses from different hypotheses and then determines weights to integrate the responses for recognizing the characters. However, determining the weight for integration is the main problem to achieve good accuracy. A MRF (Markov Random Field) based method has been proposed by Mishra et al. (Mishra, Alahari, & Jawahar, 2011) for text binarization in scene image. This method works well for camera-based images and scanned document images. For classifying the text, Wakahara and Kita 2011 (Wakahara & Kita, 2011) explored a k-means clustering. A hybrid binarization method has been proposed by

Milyaey et al.2015 (Milyaev, Barinova, Novikova, Kohli, & Lempitsky, 2015) for both plain and natural scene images. The performance of above binarization methods is satisfactory for caption but not scene. Recently Mishra 2017 (Mishra, Alahari, & Jawahar, 2017), has integrated some new features. Using character-like strokes, they extracted possible candidate text regions in an iterative scheme. Later, a stroke based method has been introduced by them with the color based energy function. Very recently, a MSER based feature detector has been proposed by Baran et al. (Baran, Partila, & Wilk, 2018) to extract connected component from image. For separating text and non-text component obtained by MSER, different oriented contour and geometrical have been applied. Finally, an OCR system is used for recognizing optimal words. All these methods need a fixed set of rules and threshold for binarization of documents.

Therefore, from the review of binarization-based methods, it can be inferred that these algorithms focus on those scene text where it has good quality, clarity, high contrast, uniform color, uniform text size, and big font. In addition, global or local thresholds govern the most of the methods. Since video suffers low contrast and resolution, the performance of existing binarization works may not be satisfactory in video.

2.6.1.3 Video Text Binarization

For overcoming the problems associated with this adaptive thresholding based approach in natural scene image, probability-based methods are exploited for video text binarization. This approach labels each pixel based on some probability measure. For increasing the rate of character recognition through binarization, some researchers applied color, stroke, edge, and corner clue. To enhance the resolution of the text image, Li et al. (Huiping Li, Kia, & Doermann, 1999) proposed an interpolation method which is named as Shannon theory. A multiple frame-based technique has been proposed by Li et al.(Huiping Li & Doermann, 1999) to isolate background noise from the foreground text. This approach works well when the movements is different between background and text. Odobez and Chen (Dobez & Chen, 2002) explored a binarization method using MRF and grayscale consistency constraint (GCC). However, the method fails to address the complex background in video. A multiple hypothesis framework has been proposed in (D. Chen, Odobez, & Bourlard, 2004) for text recognition in video. Here, for generating binary image, an edge detector is exploited. Then the method extracts gradient and texture features. However, for images with the varying background, edge detector might not perform well. Layer based binarization has been presented by Yaakov Navon (Navon, 2008) from images. Kim and Kim (W. Kim & Kim, 2009) has introduced a novel framework to binarize caption text in the video. This method assumes that transient colors exist between edited text and its adjoining background. This method works on high-quality scene image, but not low resolution video scene text. Hua et al.(Hua, Yin, & Zhang, 2002) and Yi et al. (J. Yi, Peng, & Xiao, 2009a) works on video text with complex background. Hua et al.(Hua et al., 2002), extracts text using multiple frame integration. Afterwards, it uses block-based adaptive thresholding procedure and K-means clustering. These techniques help in segregating the background and text pixel in word level for recognition. In the second method (J. Yi et al., 2009a), the text blocks are divided using the local Otsu into the text and non-text parts. This method works only on the Chinese language. Similarly, Jian et al.(J. Yi, Peng, & Xiao, 2009b) used multiple frame integration to increase recognition accuracy in complex background of video. An edge-based binarization method has been proposed by Zhou et al. (Zhiwei, Linlin, & Lim, 2010) for video text image to improve the video character recognition rate. This method performs well for small gaps. However, it cannot handle big gaps in character component. Moreover, the main target is big and caption text but not scene. A binarization method has been introduced by Ntirogiannis et al. (Ntirogiannis, Gatos, & Pratikakis, 2011) based on convex hull analysis and stroke width extraction. In this method, emphasized has been given on artificial text. A fusion approach based on wavelet sub-bands has been proposed by Roy et al. (S. Roy et al., 2012) for recognizing the image through OCR. They have used clustering across row and column on fusion image obtained from wavelet subbands and gradient direction images to obtain text candidate. To refine the performance, connected components of text candidate are analyzed. Although the method gives a satisfactory result for high resolution, big font video text image, the method fails for small fonts and low resolution video images. Moreover, the main focus is not arbitrary orientation, only horizontal direction. Shi et al. (Cunzhao Shi, Xiao, Wang, & Zhang, 2012), has explored graph cut. Instead of the using whole image, sub-divided images are considered for binarization. Based on color space clustering and MRF, Zhang and Wang (Z. Zhang & Wang, 2013a) introduced a binarization method for superimposed text in the video. In this approach, emphasize has been given in Chinese text. Using temporal information, Phan et al.(Phan, Shivakumara, Tian, & Tan, 2013) explored a recognition method using SIFT and SWT. Multilayer perceptron (MLP) based method has been proposed in (Banerjee, Bhattacharya, & Chaudhuri, 2014) to classify text. In this approach, SIFT is used on the full frame to search the text.

Most of the above methods, mentioned above compute the probability on pixel level. For this reason, the complex background cannot be handled; thus text-like background feature is generated. Moreover, high-resolution video caption text is the target of most of the binarization methods but not low-resolution scene texts in the video.

Overall, the above methods do not consider multi-type and multi-oriented text images and videos for recognition.

2.6.1.4 Multi-Type-Oriented Video Text Binarization

To increase recognition rate on multi-type-oriented video text, Wu et al. (Y. Wu et al., 2016) have restored complete character contours. They utilized directly gray values instead of choosing or optimizing threshold in video/scene images. For identification of stroke candidate pixels, the strength of zero crossing points, which are obtained by the Laplacian is explored. After identifying candidate pixels pair, a new symmetry feature has been introduced using Fourier phase angles and gradient magnitude. For tracking based text recognition, Tian et al. (S. Tian, Yin, Su, & Hao, 2018), presented an agglomerative hierarchical clustering algorithm and a temporal over-segmentation technique. A multi-frame integration is employed by the author using voting strategy for text recognition in video.

Overall, it can be said these OCR based methods are computationally inexpensive and simple; however, big font and caption text are only prioritized by these methods. The binarization approach may not preserve characters, rather it loses shapes for the images of different contrasts and background complexities. Thus these methods fail if scene texts and graphic texts are located at the same frame in the video. Moreover, binary based recognition method entails a 'cleaned' binary text image. Therefore, to get a cleaned binary image, various pre-processing and post-processing techniques have been applied before the final recognition step, namely: gray-scale filtering, segmentation using edge, boundary detection, integration using multiple frame, and graph-based noise removing. These filtering processes increase computational complexity. In addition to this, OCR can recognize only standard fixed format in the horizontal direction and very sensitive to font size and font type. Therefore, irrespective of text type, font, contrast, and orientation, improving the accuracy of video character recognition through binarization is challenging.

2.6.2 Recognition through Classifier

To prevent the problems of binarization such as loss of information, shapes, and disconnections, the methods proposed classifier for recognition rather than using available OCR which does not accept impaired shapes. This section reviews the methods which use classifier for recognition.

2.6.2.1 Document Text Recognition

There are some recognition methods by researchers who applied machine learning approaches (Babaguchi, Yamada, Kise, & Tezuka, 1991; Kuk, Cho, & Lee, 2008; Lelore & Bouchara, 2009) in case of document processing. There are different ways of using machine learning, including Markov Random Field (Lelore & Bouchara, 2009; Cunzhao Shi et al., 2012) , self-learning (Bolan, Shijian, & Tan, 2010), Gabor filters (A. K. Jain & Bhattacharjee, 1992; Sehad, Chibani, & Cheriet, 2014) and Laplacian energy (Ayyalasomayajula & Brun, 2014). There are some approaches which work on sub-regions of a document by searching appropriate binarization method. For example, Chou et al. (Chou, Lin, & Chang, 2010) partitioned an input image into sub-regions and then used SVM on those sub-regions to select the type of binarization algorithm to be used. Chamchong et al. (Chamchong & Fung, 2010) tried to opt the best binarization algorithm among a set of possible algorithms using a NN. Neural network selects the threshold according to the input histograms of image. After that, the chosen threshold binarizes the images. In another work, based on gray value of centered pixel and neighboring pixels, Sari et al. (Sari, Kefali, & Bahi, 2012) used MLPs to label a pixel as foreground or background. LSTM-based approach, which processes long sequences efficiently has been proposed by Afzal et al. (Afzal et al., 2015) for document image binarization. In this approach, a sequence of pixels represents an image. A single pixel as a background or text is classified using a 2D LSTM. LSTM works well without any feature extraction and parameter tuning. Furthermore, any pre or post-processing is not required to get semantic meaning of text. In case of non-uniform illuminated document, classifier-based binarization methods has been applied to choose a local thresholds that need to be elected according to different brightness conditions. For this purpose, Chou et al. (Chou et al.,

2010) proposed a region-based binarization for document image. From the information provided by training samples, decision functions is constructed using SVM. And finally, these rules are used to judge which binarization algorithm will be imposed for that region. A discriminative structural-based classifier has been proposed by Ahmadi et al. (Ahmadi, Azimifar, Shams, Famouri, & Shafiee, 2015). The main disadvantage of the method is that the reported accuracies are not consistent for different databases. Recently, in (Amrouch & Rabi, 2017) author has used CNN for text recognition. In (Wigington et al., 2017) author has introduced two data augmentation and profile normalization techniques, which are used with a CNN-LSTM. Very recently, a system for offline recognition cursive Arabic handwritten text has been proposed by Rabi et al. (Rabi, Amrouch, & Mahani, 2018) using HMMs. The system does not need explicit segmentation. After estimating baseline of text, statistical and geometric features are extracted to integrate the pixel distribution and text characteristics in the word image.

In summary, in case of blurred, and distorted image with small font size, the extracted features are not robust. It is true that blur is common for scene and video due to text movements and camera movements. Therefore, the performance of existing classifier-based methods is not satisfactory for scene and video.

2.6.2.2 Natural Scene Text Recognition

For the text recognition in natural scene images, a large number of works have addressed the issues associated with the document recognition classifier. These methods rely on the spatial frequency analysis, edge detection, and binarization of images. For example, a single framework has been proposed by Tu et al. (Tu, Chen, Yuille, & Zhu, 2005) for detection, segmentation, and recognition. In (Weilin Huang, Lin, Yang, & Wang, 2013), the author presented a top down-based recognition method which works in natural scene text. It extracts rotation invariant features to identify the text part in image.

Wang et al. (K. Wang & Belongie, 2010) exploited the HOG for scene text recognition. This method employs lexicons in post-processing step to increase the recognition accuracy. A text recognition framework has been presented by Gonzalez et al. (A. Gonzalez, Bergasa, & Yebes, 2014) on traffic panels. They find the interesting key points from the specific color pattern and then recognize the character in the interested part. In (Shivakumara, Raghavendra, et al., 2017), the author proposed a text detection approach using the different features, namely: Fourier, Polar descriptor, and Pseudo-Zernike moments. Finally, the author used SVM based classification for recognition. An end-to-end recognition system has been proposed by Jaderberg et al. (Jaderberg, Simonyan, Vedaldi, & Zisserman, 2014) from scene images. They used a deep CNN. Using a deep learning, they (Jaderberg, Vedaldi, & Zisserman, 2014) proposed text spotting. For word recognition, a lexicon free segmentation based approach has been introduced by Alsharif et al. (Alsharif & Pineau, 2013) in natural scene. Hybrid HMM model is employed for segmentation of words into characters. Viterbi algorithm has been used for recognition. Chattopadhyay et al. (Chattopadhyay, Reddy, & Garain, 2013a) applied appropriate binarization algorithm among a set of candidate methods. The selection of algorithm is based on complexity of the background. For this purpose, to select one of the binarization techniques, a SVM was used. However, it is difficult to make a decision of binarization methods with the classifiers since the background is very unpredictable for video images. Novikova et al. (Novikova, Barinova, Kohli, & Lempitsky, 2012), performed word recognition by calculating the maximum a posteriori. This posteriori is estimated using character appearance and the language model jointly. Mishra et al. (Mishra, Alahari, & Jawahar, 2012a) used a large dictionary to incorporate a robust language model for increasing the recognition accuracy. In another work, Mishra et al. (Mishra, Alahari, & Jawahar, 2012b) fused bottom-up and top-down cues using character and language, respectively for recognizing text. In this approach, local maximum character is detected using sliding window. The interactions between characters are jointly model using CRF. Roy et al. (S. Roy, Roy, et al., 2013a) explored HMM using some gradient related statistical features. This recognition method works for multi-oriented scene. However, they did not consider video text.

Deep learning has remarkably overcome many challenging problems in the computer vision area. Coates et al. (Coates et al., 2011) made an initial attempt using an unsupervised learning approach for character recognition. Later, a multi-layer neural network is proposed by Wang et al.(T. Wang, Wu, Coates, & Ng, 2012) adapted for text recognition. In (Weilin Huang, Qiao, & Tang, 2014), Huang *et al.* integrated MSER and CNN for detecting scene text. The MSER extracts possible text component, whereas a CNN is used to identify true text component by separating the connections of multiple components. Contemporary to this method, a recurrent neural network (RNN) framework has been proposed by Su and Lu (Su & Lu, 2014) for recognizing scene text. Initially, a sequential histogram of gradient features is extracted from word image. Finally, a RNN is applied to classify these features into words. He et al. (P. He, Huang, Qiao, Loy, & Tang, 2016) proposed a deep recurrent model, building on LSTM, to robustly recognize the generated CNN feature from scene text. Recently, Cheng et al. proposed (Cheng et al., 2017) a Text Attentional CNN (Text-CNN). Here, the mask of a text region, the binary information of the text/non-text, and ground-truth of character are used to train the network. This rich supervised information augment the robustness against the noisy and complex background. A novel Convolutional Recurrent NN (CRNN), has been explored by Shi et al. (B. Shi, Bai, & Yao, 2017). This fuses the advantages of both CNN and RNN for scene text recognition.

Although the method described above gives satisfactory result for high resolution, big font scene image, but the method fails for small fonts and low contrast text, which are common for video images.

2.6.2.3 Video Text Recognition

There are some methods which explore classifier for recognition of video text. For example, Saidane and Garcia (Saidane & Garcia, 2007) proposed a binarization method using CNN in case of scene/video image. The number of instances used in training governs the accuracy. Roy et al.(S. Roy, Roy, et al., 2013b) proposed a method involving two level of word recognition/verification using HMM and CNN in scene/video images. A well-known feature descriptors, namely, SIFT, HOG, SURF, and LBP are extracted. The features extracted based on descriptors work well for high contrast images and horizontal video image. An unconstrained approach has been proposed by Jain et al. (M. Jain, Mathew, & Jawahar, 2017) for scene text and video text recognition. However, the method is limited to Arabic script. A 3-gram language model (LM) has been proposed by Elagouni et al. (Elagouni, Garcia, Mamalet, & Sébillot, 2014) to enhance the text recognition in videos. This approach consisted of two steps: the first step segregates texts into isolated characters and the next step combines window classification results obtained from a graph model and multi-scale scanning window. To find the possible path or word, joint probabilities is estimated using LM. Generally, LM is applied to correct some errors pertaining to the ambiguous text character, and over or under segmentation. In (Rong et al., 2014), the author proposed a video text recognition algorithm using two level fusion approaches, mainly majority voting model and CRF. The voting model predicts labels of scene text character (STC) in all frames and the CRF model integrates STC prediction scores from multi-frame using vocabulary. Similarly, Greenhalgh and Mirmehdi (Greenhalgh & Mirmehdi, 2015) combined the text results from temporal frames. Initially, individual OCR results are compared based on text size between frames. Then, the recognition labels are combined from the result of ten frames. After that from each tracked word, a histogram of OCR is constructed. The recognition confidence governs the weight of histogram. Finally, the highest value obtained from histogram result decides the recognized word for each frame. Yousfi &Garcia (Yousfi, Berrani, & Garcia, 2015) has extracted a sequences of features from Arabic text. Using the BLSTM-CTC schema, text recognition is performed. Recently, to improve LSTM in video text recognition, Yousfi et al. (Yousfi, Berrani, & Garcia, 2017) designed recurrent connectionist language modeling based on Arabic text. A simple RNN model has been used in this work. Language models are learned using a Maximum Entropy language model. Wang et al.(Xiaobing Wang et al., 2017) has used the deep neural network to recognize text in the video. An ensemble of CNN's trained on synthetic data is adapted for detecting and recognizing characters by Xu 2018 (Y. Xu et al., 2018). This method is limited to the East Asian character.

It is observed from the above discussion that all methods concentrate on recognizing specific script but not multi-script. Furthermore, the window size cannot be fixed for arbitrarily-oriented characters for extracting features; hence the insufficient information obtained from character leads to poor performance. For deep learning set, as variations on dataset increases, fixing the values of parameters also increases. As a result, there is a huge difference between the accuracies on different datasets. Therefore, for the multi-type images, the same method discussed in above might not perform well.

2.6.2.4 Multi-Type-Oriented Video Text Recognition

For recognizing multi-oriented texts, there exist some approaches (Umapada Pal, Roy, Tripathy, & Lladós, 2010; P. P. Roy, Pal, Lladós, & Delalandre, 2012) in graphical documents. An approach for recognizing the multi-scaled and multi-oriented character has been proposed by Pal et al. (Umapada Pal et al., 2010). However, the target of this method is scanned images; thus performance is not good enough in case of natural scene or video. A fusion of HMM and CNN is proposed in (S. Roy, Roy, Shivakumara, & Pal, 2013) to achieve good recognition rate. To find character alignment of a word, sequential gradient features are extracted with HMM, and in the later stage, the character alignments are verified by CNN. Recently, a novel multi-type text recognition approach has been proposed in video by Bhunia et al.(Bhunia, Kumar, Roy, Balasubramanian, & Pal, 2017) based on color channel selection. The approach consisted of three steps. Initially, a color channel is automatically chosen. A sliding window moves over selected color channel to extract PHOG feature. Then extracted features are processed using HMM. Finally, for obtaining best recognition accuracy, a multi-label SVM classifier is used. Recently, a hybrid CNN-RNN network with a Connectionist Temporal Classification has been proposed for multi-type oriented text recognition by Jain et al. (M. Jain et al., 2017). But this is limited to Arabic text. An adaptive ensemble of deep neural networks (AdaDNNs), is proposed to select and adaptively combine classifier components at different iterations from the whole learning system for recognizing multi-type and multi-oriented text. Furthermore, the ensemble is formulated as a Bayesian framework for classifier weighting and combination Chun et al. (C. Yang, Yin, Li, et al., 2017).

In summary, most of the above recognition approaches explore well-known descriptors, namely, SIFT and HOG for feature extraction. Then they use classifiers with a large number of lexicons for recognizing texts in images. A major shortcoming is that these methods are restricted to a predefined lexicon. The association between training data and lexicons depends on the script. As a result, this kind of approach restricts the possibility of flexibility of working on different scripts without updating the training scheme and performance of the method drops. Moreover, the classifier is expected to output high probability for test sample if similar kind of sample text image is considered in training phase. As these algorithms need sufficiently a huge amount of data during the training phase, therefore extracting features from each sample are computationally expensive. Most of the methods are not evaluated on words in video and require high contrast for scene images. From the review of existing works, it is noted that majority of the algorithms give a satisfactory result for the particular data type, on which the methods concentrate. But none of the methods deal more than two types of data for recognition. Besides, multi-script recognition in case of Indian scripts is still at the blooming stage. Therefore, these approaches may not cope with the new problems of multi-type and multi-oriented video text recognition. Therefore, there is a scope for developing a robust recognition algorithm, particularly for different orientations and variant type of text in video and scene image.

2.7 Summary

Overall, the approaches which are proposed for video text classification, Text enhancement, text extraction, text type classification and text recognition are reviewed. The fact of working with recognition of video text raises several problems to tackle. By reviewing the existing methods, it is noted that there are still major problems as discussed in below for text recognition in video.

(1) It is noted from classification literature that these methods achieve good results for a particular video type but show lower performances for heterogeneous types of videos. Therefore, there is a huge difference between the accuracies on different datasets. Thus, there is an immense scope for the automated classification of different video text categories before choosing an appropriate text detection and recognition method to achieve better results.

(2) The existing enhancement methods enhance the video text frame, but they introduce too many noises. From the discussion of the literature survey, it can be inferred that there is no consistent enhancement method for reducing the effect of

Laplacian operation. Till to date, there is not generalized enhancement model for the distortions caused by Laplacian or gradient so far.

(3) Though there are many conventional text detection, they perform well for text lines; thus these methods do not yield good result on bib numbers. Most of the existing approaches consider text word to be straight in image and rarely search curvy-linear or non-horizontal or arbitrary word due to its complexity. So, the problem of detecting word irrespective of orientation did not receive a great deal of attention in the literature. In addition, the spotting approach is limited to video images affected by low contrast and complex background but not natural scene images and license plate images. Therefore, multi-oriented bib detection/spotting entails a special attention in such video or images to increase recognition accuracy.

(4) It is confirmed from text type classification that the performances of the existing methods degrade and are unsatisfactory due to the presence of caption and scene text types in video frames. Though, temporal frames are explored by some approaches, the determination of the required frames is tricky. Generally, based on experiments, the methods generally choose the required number of frames. The threshold chosen heuristically might not perform well for variant datasets and conditions. Therefore, achieving better results for video with a single method is not as easy as for those images having only one type of text, such as natural scene images and document images.

(5) Most of the researcher focus their attention only on binary image for recognition and discard the gray part which could be obtained from color image. So, there is immense scope of exploring gray value without the need of binarization. In addition, the classifier-based method needs lots of training data with a vast lexicons for classifying non-text and text pixels. Traditionally, recognition approaches concentrate on horizontal, slightly non-horizontal and mostly on

single type of text. Moreover, the association between data and lexicons depends on the script. As a result, this kind of approach restricts the possibility of flexibility of working in different script. Therefore, the methods fail in recognizing multi-type, multi-oriented and multi-lingual text in video.

The above shortcomings of the current approaches have motivated to propose new methods, which are able to recognize multi-oriented and multi-type text (caption and scene) in videos accurately and efficiently, irrespective of font type, size, and scripts.

CHAPTER 3: FUZZY-ROUGH BASED IMAGE VIDEO CATEGORIZATION

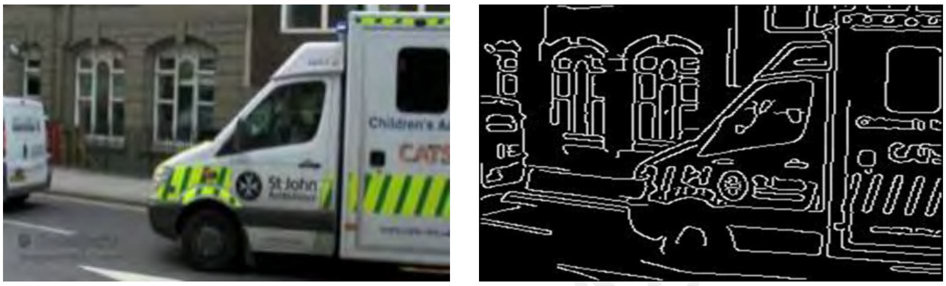
3.1 Background

It is noted from the review presented in the previous chapter on video classification, video enhancement, video text detection/spotting, video text type classification, and video text recognition that robust method for video type classification, enhancement, text detection and recognition is essential for achieving better recognition accuracy on multi-type text recognition in multi-oriented environment.

This chapter presents a new approach for categorizing scene videos into different classes, namely, Animation, Outlet, Sports, e-Learning, Medical, Weather, Defense, Economics, Animal Planet, and Technology, to improve text detection and recognition performance. For this purpose, Fuzzy concept is explored in two different way, namely fuzzy-mass based approach and fuzzy-rough based approach for video type classification.

3.2 Fuzzy-Mass based Approach for Video Type Classification

For the edge images of videos, such as Defense, Sports, and Medical, it can be observed that the degree of cursiveness is more compared to the edge images of Economics and Weather videos. This is due to the fact that the backgrounds of the ocean, sky and forest in Defense video, stadium and ground with greenery in Sports video, and buildings/streets in Medical video usually produce dense curved edges compared to the backgrounds in Economics and Weather videos, where less cluttered backgrounds can be seen. One such example can be seen in Figure 3.1, where the medical video frame contains more curved edge components than straight edge components. In this work, canny edge operator is chosen as it gives fine details of edges for low contrast as well as high contrast images compared to Sobel, which usually gives fine details only for high contrast images (Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010). Besides, edge image helps in extracting features at the component level which reduces the number of computations. It is illustrated in Figure 3.1, where (a) denotes a sample medical video frame and (b) denotes its canny edge image. However, at individual edge components, some parts of edge components appear straight, and another part of components appear to curve. This leads to uncertainty in identifying exact edge components as straight, and curved. To overcome this uncertainty, a new Fuzzy based approach along with mass and a spatial feature has been proposed for classification.



(a) Input Medical video frame (b) Canny edge image

Figure 3.1: Examples of dense curved edges for Medical video frame.

For classified straight and curved edge components, the proposed approach extracts features to study the degree of the mass of different video types both locally and globally. Then the proposed approach combines both local and global features to achieve good classification rate. To extract local features, a successive ellipse over edge image has been drawn rather than defining a window because ellipse helps in defining granule to extract mass based features (Ting, Zhou, Liu, & Tan, 2013). For global features, the proposed approach uses the whole edge image without ellipse drawing. The framework of the proposed approach is shown in Figure 3.2.

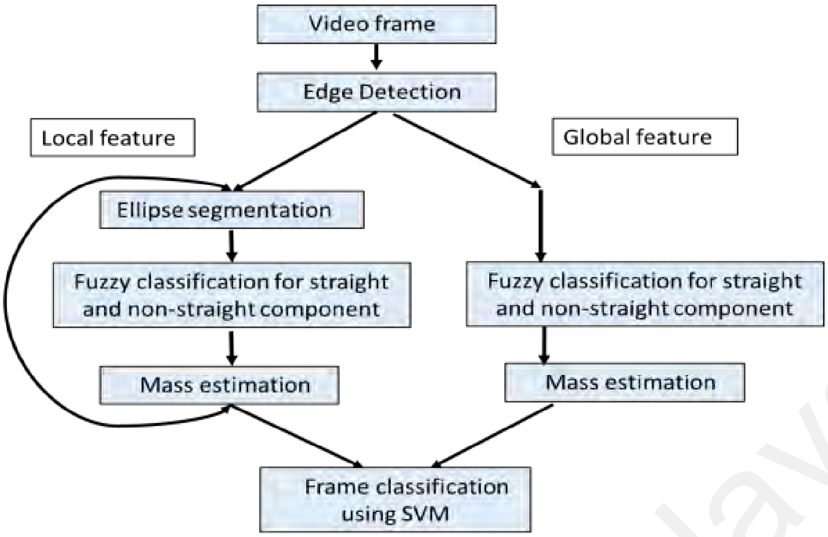
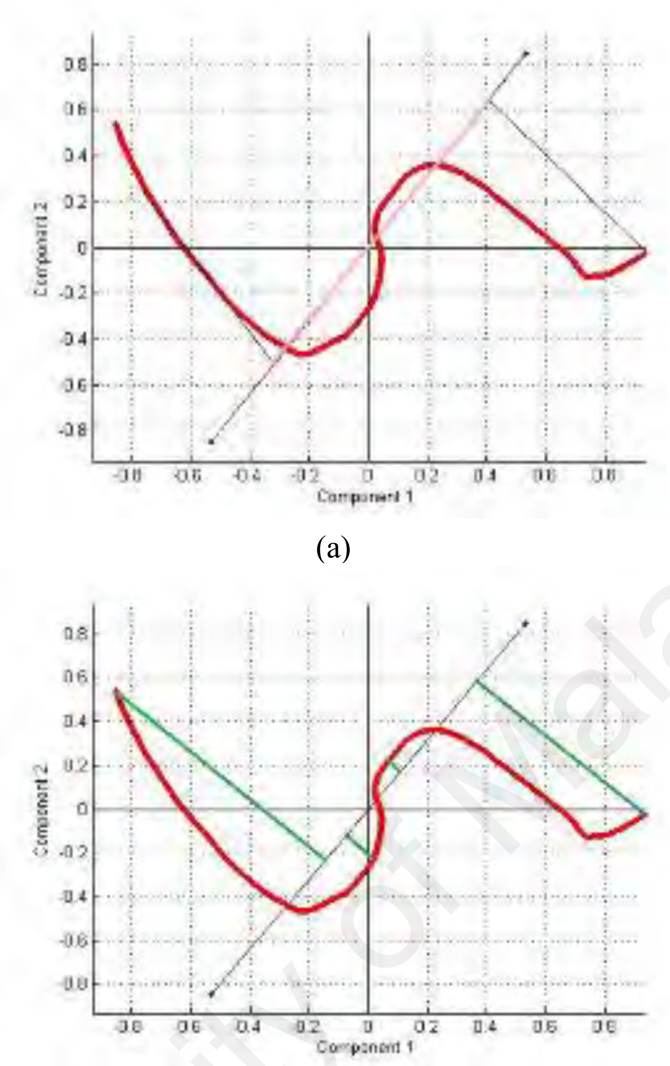

Video frame
Edge Detection
Local feature
Global feature
Ellipse segmentation
Fuzzy classification for straight and non-straight component
Fuzzy classification for straight and non-straight component
Mass estimation
Mass estimation
Frame classification using SVM

Figure 3.2: Framework of the proposed approach.

3.2.1 Straight and Curved Edge Components Classification

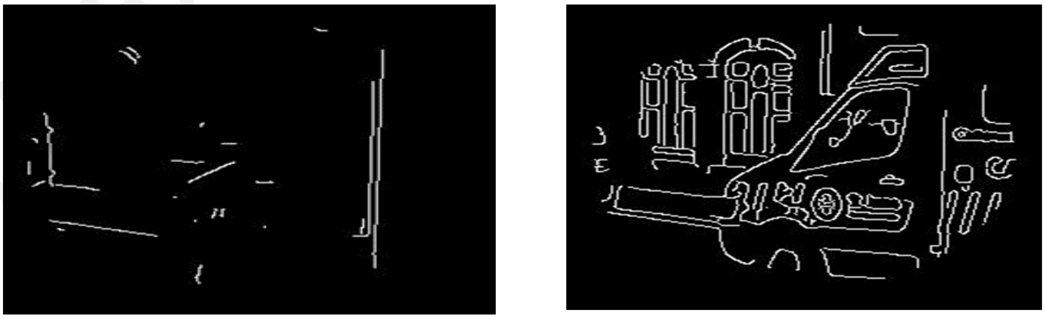
For each edge component, the proposed approach analyzes its area under the curve to classify it as straight or curved. This is valid because if an edge component is curved, it is expected that a larger area under the curve or almost zero area under straight. Therefore, the proposed approach performs projection onto principal axis from coordinates of the edge component. The sum of these projections results in an approximated area under the curve. To make the area under the curve scale invariant, the estimated area is divided by the length of the segment of the principal axis under the curve. The sum of the square of perpendicular distances to the principal component axis is used as an error metric for choosing the best straight line that fit data (Nogueira et al., 2017).



(a)

(b)

Figure 3.3: Examples of Fuzzy membership functions for classifying straight and curved edge components.



(a) Straight components

(b) Curved components

Figure 3.4: Results of Fuzzy logic based straight and curved edge components classification of the edge image shown in Figure 3.1.

Now the question is how to fix an automatic threshold for the area estimated under straight and curved edges. Fuzzy logic based on area approximation has been introduced to tackle the issue. A principal component axis for each edge component is plotted as shown in Figure 3.3, where it can be seen that the principal axis is given by principal component analysis and its projections. The value of the area obtained by the above procedure can lie in the region $[0, \infty)$. Then scaled Gaussian function has been proposed on $[0, \infty)$ interval for finding the fuzzy membership function as $y = e^{-(x^2/2)}$, which has value $y = 1$ at $x = 0$. The reason to choose Gaussian function is that it is a smooth function defined for all the values greater than or equal to 0. Further, centroid method has been used for defuzzification (G. Zhao & Pietikainen, 2007).

For finding a centroid, note that $\int f(x)x\,dx = \int_0^\infty e^{-(x^2/2)}\,x\,dx = 1$. A centroid is given by $\frac{\int f(x)\,x\,dx}{\int f(x)\,dx}$. The denominator is the area under a curve. The area under standardized Gaussian is known to be 1. So, $\int_0^\infty e^{-(x^2/2)}\,x\,dx = \sqrt{}(\pi/2)$. Therefore, the required value of x at centroid is $\sqrt{\frac{2}{\pi}}$. This is the required threshold. The example shown in Figure 3.3 is classified as a curved edge component because it gives a value larger than the threshold given by Fuzzy logic. Sample classified straight and curved edge components can be seen in Figure 3.4 for the edge image in Figure 3.1 (b).

3.2.2 Mass-based Features Extraction from Edge Components for Video Image Categorization

For the classified straight and curved edge components in the previous section, dense observation has been extracted as discussed in the earlier part of this section by estimating mass rather than density because mass is estimated based on small granularity with its elements. Hence, mass estimation is simple, efficient and robust compared to density estimation (Ting et al., 2013). Mass (m) is generally defined as the number of the points in a region, i.e., $m = \sum_{i=1}^{N} p_i$, where p denotes an individual component in the region, and N corresponds to the total count of straight and curve components in the image. Here a region is granularly defined by an ellipse.

In order to extract local information, consecutive ellipses are drawn as shown in Figure 3.5 (a), where wave patterns can be seen given by consecutive ellipse formation. The step size to draw consecutive ellipses has been estimated depending on the major axis of the smallest edge component present in the current ellipse. The process of consecutive drawing ellipse stops when an ellipse gets at least two edge components. Each ellipse is considered as a granule for local features extraction in this work. For each ellipse, the proposed approach first gets straight and curved edge components using the method in Section 3.2.1 and then extracts features separately. Next, the proposed approach considers the sum of all the feature values extracted respectively from each ellipse. The sum is divided by the total edge component. This generates two mass based features separately for straight and curved groups.

Therefore, the features f_1 and f_2 can be defined as equation (3.1) and equation (3.2), where m_s denotes the mass from straight edge components, and m_c denotes curved edges components. The sum is divided by the total number of components (*N*) in the whole image (edge image).

$$f_1 = \frac{m_{st}}{N} \qquad (3.1), \qquad\qquad f_2 = \frac{m_{cu}}{N} \qquad (3.2)$$

Similarly, the presented approach extracts local features using spatial proximity, computed between the centroids of straight and curved edge components in respective straight and curved component clusters, and takes standard deviation for the proximity matrices, separately, which can be defined as f_3 and f_4 in equation (3.3) and equation (3.4), respectively. Finally, the sums of standard deviations of all the ellipses divided by the total number of edge components in the edge image considered as features. The centroids for straight components and curved edge components can be seen in Figure 3.5 (b) and Figure 3.5 (c), respectively.

$$f_3 = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} d_i^{st} - \overline{d^{st}})^2} \qquad (3.3), \quad f_4 = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} d_i^{cu} - \overline{d^{cu}})^2} \qquad (3.4)$$

where d^{st} denotes the distance between straight components, and d^{cu} denotes the distance among curvy components.

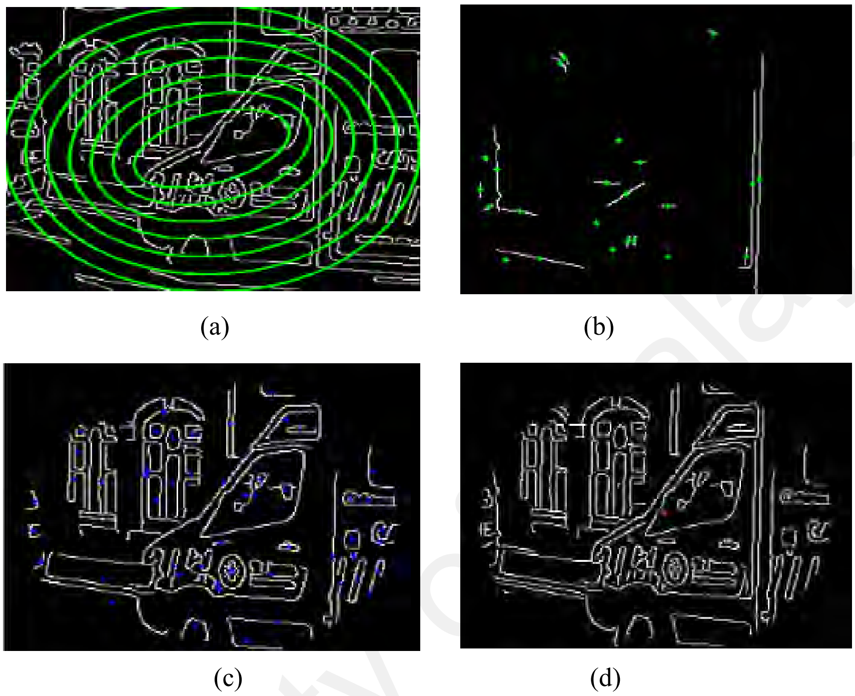


Figure 3.5: Local feature extraction using mass estimation and proximity between centroids and straight-curved edge components: (a) Consecutive ellipse to extract local information, (b) Denotes centroids for classified straight edge components, (c) Denotes centroids.

The distance between image centroid and straight edge components centroid is calculated, say f_5; image centroid and curved edge components centroid, say, f_6; and straight edge components centroid and curved edge components centroid, say, f_7, respectively. The sample centroids for the whole image, the straight edge components, and the curved edge components can be seen in Figure 3.5 (d), where red, green and blue pixels represent image centroid, straight edge components centroid and curved edge components centroid, respectively. As mentioned above, the final features are obtained by summing up the feature values of all the ellipses and being divided by *N*. This results in seven local features. In the same way, the similar seven features are extracted for the whole edge image without drawing ellipses, which are considered as global features. In total, fourteen features are obtained for classification in this work. Inspired by the work presented in (G. Zhao & Pietikainen, 2007) for facial expression recognition, where SVM has been successfully used, a similar procedure has been followed for video text frames classification in this work by feeding the 14 extracted features to an SVM classifier. As suggested in (G. Zhao & Pietikainen, 2007) where an SVM is used for discriminating two sets of data to achieve good accuracy for facial expression recognition, the five-class video classification problem is partitioned into a 10 class problem to achieve the same for classification. For example, defense-economics, defense-sports, defense-medical, defense-weather, economics-sports without any repetition of groups. Then a voting scheme is used for the final classification. However, this voting may lead to having the same number of votes for more than one class. To avoid this problem, the first class appeared in the resultant voting scheme has been chosen based on experimental observations. A polynomial kernel function has been used for SVM classifier in this work. In addition, for training and testing, 10 fold cross-validation procedure has been followed. The final confusion matrix is obtained after averaging 10 confusion matrices.

In summary, the proposed approach is limited to simple and few classes of video types but not complex classes. To overcome this limitation, to deal with this ambiguity and uncertainty between the classes, a Rough-Fuzzy based approach has been proposed.

3.3 Rough-Fuzzy based Approach

In this work, a unified system is proposed as shown in Figure 3.6, where the classification step identifies video types, the text detection step extracts text lines from video frames, the binarization step extracts foreground from text lines, and finally, OCR recognizes texts. Figure 3.6 shows the parameters, which are derived automatically with the help of training samples for respective steps. Since the main goal of this work is video categorization, the focus is given to the classification method and use text detection and recognition performances for validating the proposed classification.

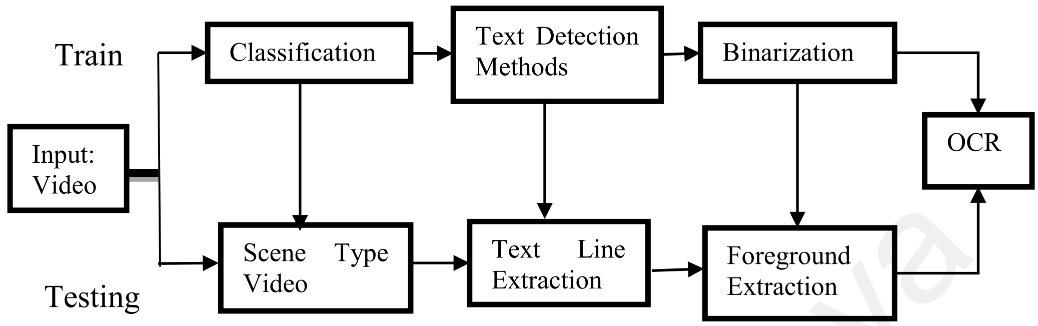
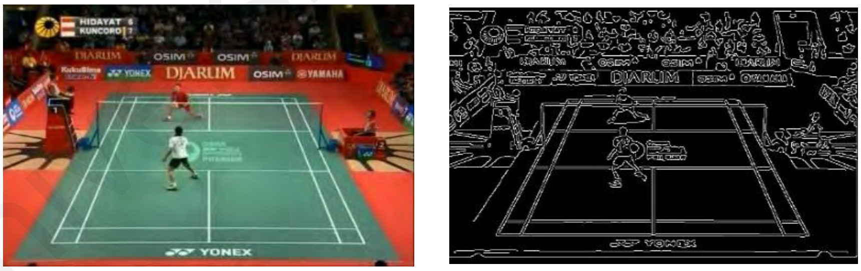


Figure 3.6: Unified frame work of the proposed method.

For a given scene type video, in this work, key frames and neighboring temporal frames are extracted for classifying scene type videos. The Canny operator is applied to each frame. This operator gives edge components with their structures (L. Wu et al., 2015). The advantage is that it saves a large number of computations by extracting features at the component level instead of pixel level. For classifying each video frame as a particular class according to the nature of its content, a new combination of rough set and fuzzy logic has been proposed to group edge components as Line, Rectangle, Square, Parallelogram, Circle, Loop, Ellipse and Trapezium based on geometric shapes of edge components to extract local information of the frame. Since the considered video categorization problem is complex, one can expect uncertainty in defining shapes of edge components. Therefore, to deal such situations, the rough set is introduced to estimate lower and upper boundary approximations (Yu, Pedrycz, & Miao, 2013), which gives boundaries for extracting shapes of edge components. Due to the foreground and background variations, an approximated shape by rough set may overlap with other shapes of edge components. This leads to confusion or uncertainty. Therefore, motivated by (Fonseca & Jorge, 2000) where it is shown that fuzzy logic for recognizing elementary geometric shapes is useful in improving object recognition, fuzzy logic is introduced to recognize geometric shapes (Fonseca & Jorge, 2000; Ghanei & Faez, 2016). This step outputs eight groups according to the shapes defined by rough set and fuzzy combination for the given video frame. The eight groups are empirically determined by studying shapes of edge components for different classes.

It is true that gradient directions of edge pixels represent stroke direction distribution, which in turn provide a vital clue for extracting shapes of edge components (M.-K. Zhou, Zhang, Yin, & Liu, 2016). Therefore, each group is divided into several planes according to the gradient direction of pixels of edge components in each group. For each plane, further correlation and covariance features are extracted using gradient values to encode statistical and spatial correlation between stroke directions. Features are extracted for all the eight groups by this way. Furthermore, to add stability to these features, temporal frames have been explored. Finally, the feature matrix is passed to a neural network classifier for frame classification.



(a) Sports frame (b) Canny edge image

Figure 3.7: Edge component detection for the sample sports input frame.

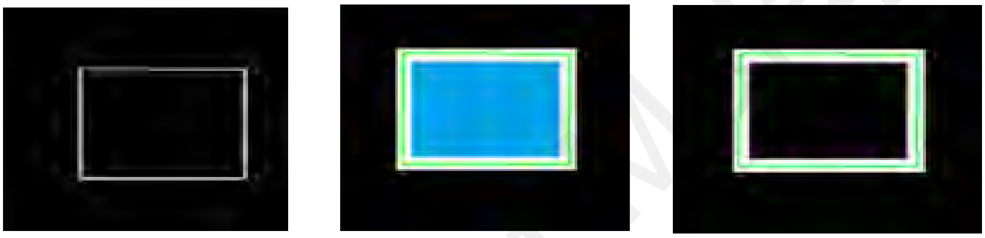
3.3.1 Grouping Edge Components based on Shapes

For a sample frame selected from sports video as shown in Figure 3.7 (a), the Canny operator is applied. The Canny operator generates an image as shown in Figure 3.7 (b), where edge components preserve the structure of components. It is also observed from

Figure 3.7 (b) that the edge components which represent background and foreground (text) have different shapes, such as rectangles, loops, ellipses, parallelograms, circles, trapeziums, squares, and lines. Among different shapes, rectangles and lines are more prominent due to the presence of courts. This observation leads to propose a new method for classifying those components into several groups to find the relationship among them for video classification.

As discussed in the earlier part of this section, the combination of rough set and fuzzy logic is extracted for grouping edge components of each frame of each class into different geometric shapes, namely, Rectangle, Parallelogram, Trapezium, Circle, Ellipse, Loop, and Line. The identification of Line is done using the projection of pixels on the principal component axis, while the identification as Loop is done by checking whether a component is split into several sub-components after removing a few pixels. Since Square is a special case of the rectangle, the same rectangle steps are used for the identification of Square. The identifications of Rectangle, Parallelogram, Trapezium, Circle, and Ellipse are done as follows: For each edge component C as shape S, the smallest object of shape S which covers C (named it the mask of component C and denote it by M_{CS}) is constructed. It can be expected that this mask precisely overlaps C for an ideal shape of edge component. However, this is not always true for real edge components due to irregular shapes and disconnections, where uncertainties are expected. Therefore, it is assumed, to match the component with the boundary of another set R_{CS}, such that R_{CS} is an approximation of M_{CS}. Approximation using rough set allows to ignore small errors and decide whether the component under consideration belongs to a given class *S*. However, even if this component does not belong to *S*, it is needed to decide how similar it is to class *S.* This is done by using the proposed fuzzy membership function discussed below. Figure 3.8 (a) shows an ideal component. In Fig. Figure 3.8 (b) the boundary of rough set R_{CS} is shown with white color, and the interior of R_{CS} is shown with blue color.

Note that the component is marked in green color. Fig. Figure 3.8 (c) shows the precise overlap of R_{CS} boundary and component boundary. The advantage of the combination of rough approximation and fuzzy membership to define irregular shapes of edge components can be seen in Figure 3.9, where for the component in (a), it can be seen that its boundary and interior of R_{CS} in (b). Note that the overlap between R_{CS} and the component as shown in (c) is partial for this component. From now, this approximation R_{CS} is defined as the mask. The boundary of this mask is defined using rough set theory as follows.



(a) Edge component (b) Granulation (c) Approximation

Figure 3.8: Illustrating rough approximation for an ideal edge component, where the component and its mask boundary overlap completely. (a) represents the edge component, (b) represents its mask boundary estimated as white region, while the interior of the mask is s.

Formally, the rough set is defined with lower and upper boundaries approximation as follows. Let X be the reference set ($X \subset U$ is the reference set, i.e., the set which is needed to approximate, while U is the universe and refers to all the pixels in the image). Lower approximation of X is the region where all the data definitely belong to X. Upper approximation is the region such that no point outside this region belongs to X. The difference between upper and lower approximations is defined as the boundary of the rough set. This is the region where some points belong to X, and some do not. It is illustrated in Figure 3.9, where (a) gives the sample edge component, (b) shows the lower approximation (blue color) and the upper approximation (white color + blue color), and

(c) is the result of actual overlap between the estimated rough set boundary and the edge component boundary. Formal definitions of both approximation and boundary region are given by (Pawlak & Skowron, 2007) as follows:

$$\begin{aligned}
&\textit{R\_lower approximation of X}\ \ R_*(x) = \bigcup_{x\in U}\{R(x): R(x) \subseteq X\} \\
&\textit{R\_upper approximation of X}\ R^*(x) = \bigcup_{x\in U}\{R(x): R(x)\ \cap X \neq \emptyset\} \qquad (3.5) \\
&\textit{R\_boundary of X}\ RN_R(x) = R^*(x) - \ R_*(x)
\end{aligned}$$

Boundary R_{CS} is defined as the set of all the points p such that the neighborhood of p contains some points belonging to mask M_{CS} and some points belonging to $M_{CS}{}'$ as stated in equation (3.6).

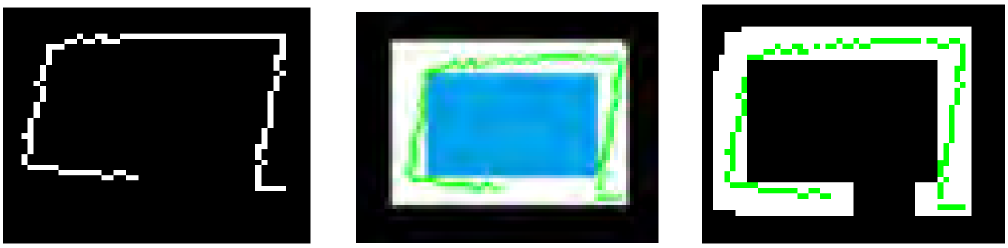
$$boundary(R_{CS}) = \{p: N(p) \cap\ M_{CS} \neq\ \emptyset \text{ and } N(p) \cap M_{CS}{}'\ \neq\ \emptyset\} \qquad (3.6)$$

Here, the lower approximation is the set of all the pixels which definitely belong to the estimated shape. This is the interior of the estimated component as shown in Figure 3.9 (b). The upper approximation is the set of all the pixels which may belong to the estimated shape. This region is the complement of the exterior of the component. The boundary is the region which is close to both interior and exterior regions of the estimated shape. Rough sets allow identifying a component to be of a perfect shape if interior and exterior regions match perfectly. This is done by using a thicker boundary region rather than the actual outline of the component for comparison as shown in Figure 3.9 (c). To estimate boundary approximation for edge component *C*, the proposed method checks whether all the pixels lie in $boundary(R_{CS})$ and those pixels which are the neighbors of the component. If a component satisfies both the conditions, it is considered as the component roughly like shape *S*.

If $boundary\ (R_{CS})$ overlaps with component C completely, it is said to be the component that is exactly like *S* as shown in Figure 3.8. Otherwise, to estimate the degree of overlap information is needed to find the closeness between the boundary and the component as shown in Figure 3.9. Let the ratio of component pixels close to masking boundary and the total number of component pixels be u_c, and the ratio of mask boundary pixels close to component pixels and the total number of mask boundary pixels be u_m. In order to find the final value which indicates how close *C* is to shape *S*, a *Z* shaped fuzzy membership function is applied to $1 - min(u_c, u_m)$ as defined in (Pawlak & Skowron, 2007). *Z* shape fuzzy membership function is a spline based one as defined in equation (3.7) and illustrated in Figure 3.10 (a).

$$
\begin{aligned}
z(x,s,t) &= 1 \; if \; x \; \leq s \\
&= 1 - 2\left(\frac{x-s}{t-s}\right)^2 \;, if \; s \leq x \; \leq \frac{s+t}{2} \\
&= 2\left(\frac{x-t}{t-s}\right)^2 \;, if \; \frac{s+t}{2} \leq x \leq t \\
&= 0 \; if \; t < \; x
\end{aligned}
\tag{3.7}
$$

If the membership function gives 1, *C* is an ideal example of the shape *S* (Lamba, 2008). Here the value is set as 1 for all the values which are less than *s*, and 0 for all the values which are greater than *t* irrespectively. However, the values always lie between *s* and *t* for real edge components. In this work, the values for *s* and *t* are experimentally determined according to the defined geometrical shapes.



(a) Edge component (b) Granulation (c) Approximation

Figure 3.9: Rough set is defined for the edge component of the sports frame where edge component boundary and mask boundary does not match completely. (a) is edge component with loss of information, (b) shows mask boundary estimated for the component as white region.

The parameters (a and b) are required to apply rough set and fuzzy logic for recognizing shapes of different edge components according to groups as follows:

Rectangle: Let the height and width of the rectangle bounding box be $2a$ and $2b$, respectively, and the centroid be (x_0, y_0). Thus *Rectangle* can be defined as in equation (3.8):

$$y \geq y_0 - a, y \leq y_0 + a, x \geq x_0 - b, x \leq x_0 + b \qquad (3.8)$$

Square: If the given component is detected to be a rectangle, check whether both the sides of the bounding box are nearly equal.

Parallelogram: A right tilted parallelogram will be defined as in equation (3.9):

$$\begin{gathered} y \leq \tan\theta\,(x - x_0 + b) + y_0 - a, \\ y \geq \tan\theta\,(x - x_0 - b) + y_0 + a, \\ y \geq y - a, y \leq y_0 + a \end{gathered} \qquad (3.9)$$

Similarly, a left-tilted parallelogram is defined as in equation (3.10):

$$\begin{gathered} y \geq -\tan\theta\,(x - x_0 + b) + y_0 + a, \\ y \leq -\tan\theta\,(x - x_0 - b) + y_0 - a, x \geq x_0 - a, \end{gathered} \qquad (3.10)$$

$$y \geq y - a, y \leq y_0 + a$$

Regular trapezium: A regular trapezium can be defined as in equation (3.11):

$$y \leq \tan\theta\,(x - x_0 + b) + y_0 - a, y \leq -\tan\theta\,(x - x_0 - b) + y_0 - a,$$

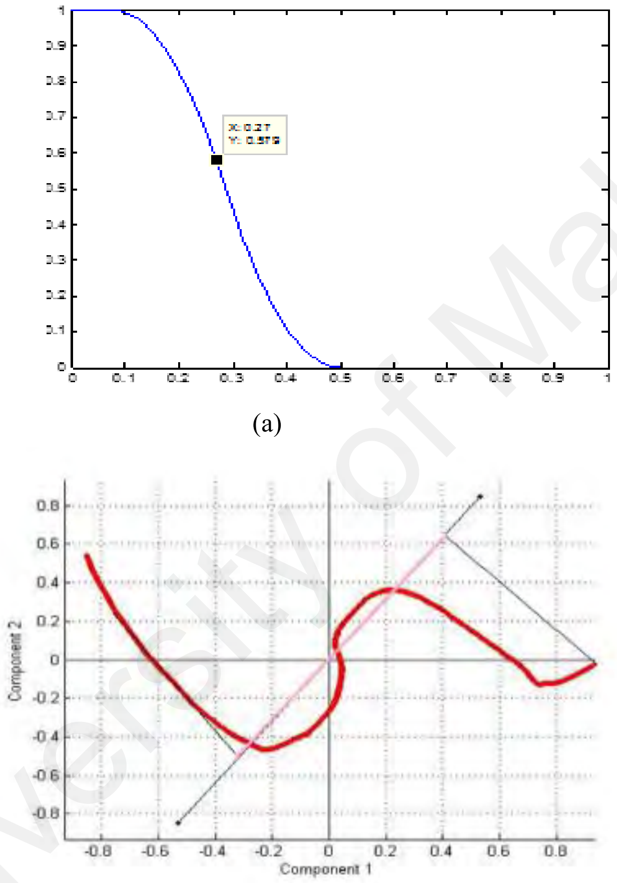
$$y \geq y - a, y \leq y_0 + a \tag{3.11}$$

Circle: The proposed method finds the centroid and the pixel which is the farthest from the centroid. The distance between the centroid and this pixel gives radius r, and centroid $(x_{0,}y_0)$ gives the center of the edge component. Therefore, the mask given by $(x - x_0)^2 + (y - y_0)^2 \leq r^2$ can be calculated using these parameters. The points on this mask can be generated as $(x_0 + r\cos\theta, y_0 + \ r\ sin\theta)$ for varying values of θ from 0 to 2π.

Ellipse: Given length $2a$ and width $2b$ and the location of the centroid (x_0, y_0), an ellipse defined by $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} \leq 1$ gives the required ellipse. The points on this ellipse can be generated as $(x_0 + a\cos\theta, y_0 + \ b\ sin\theta)$ for varying values of θ from 0 to 2π.

Loop: Suppose the given component contains *n* pixels, *n*/10 equidistant pixels are chosen from the component. For each pixel, those pixels which are at a distance of two pixels or less from the chosen pixel are removed. After removing these pixels, if the remaining edge component is still a connected component, then the chosen pixel is a part of a closed loop. Otherwise, it is a part of an open component. Repeat this procedure *n*/10 times, each time on the original image of the component. As long as the resultant edge component remains a connected component, the proposed method estimates the percentage of pixels which are a part of a loop. Note that the mask algorithm and Z shaped analysis are not used for loop test.

Line: The principal component axis for an edge component is plotted as shown in. Figure 3.10 (b), where it can be seen that the principal axis is given by principal component analysis and its projections. The value of area obtained is scaled by the square of the length of the principal axis segment corresponding to it. The value of the area obtained by the above procedure can lie in the interval$[\mathbf{0}, \infty)$. The values obtained are fed to Z shape fuzzy membership function.



(a)

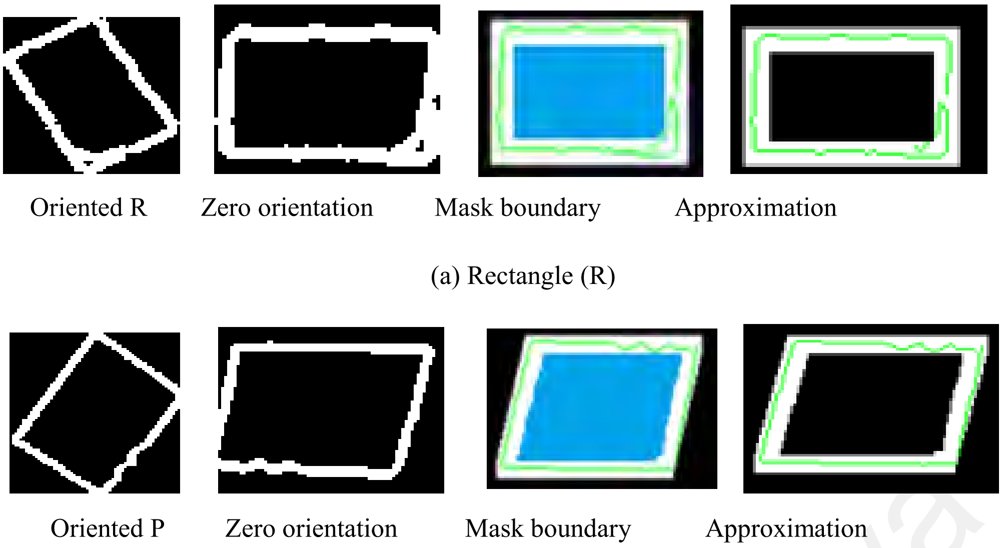
(b)

Figure 3.10: Fuzzy membership functions for classification of edge components according to shapes. (a) Z Fuzzy membership function for classification of edge components according to shapes. X axis denotes the original value calculated and Y axis shows resulting m.

The values obtained from the mask overlap algorithm and the line function are fed to Z shape fuzzy membership function. Figure 3.10 (a) shows this function with parameter values *a*=0.1 and *b*=0.5. The values of parameters *a* and *b* used for the fuzzy filter are *s* =

0.075 and *t* = 0.5 for each of rectangle, parallelogram, trapezium, circle, and ellipse. For line *s*=1e-3 and *t*=0.02.

The above process of recognizing shapes by fixing mask boundary and estimating parameters of edge components work well for those edge components having zero degree orientations with horizontal. However, in case of irregular shaped edge components, one cannot expect all the time edge components without any tilt or orientation. To overcome this problem, Principal Component Analysis (PCA) is proposed for estimating angles of edge components with respect to the X-axis. Edge components are then rotated by this angle, which results in edge components with zero orientation. In other words, PCA is proposed to use to check the orientation of each input edge component before applying the boundary approximation for them. It is true that when an edge component is rotated by an angle, little distortion is expected. However, the proposed rough set and fuzzy combination takes care of such tiny distortion affected by rotation conversion. Orientation checking using PCA is good for those shapes like Rectangle, Parallelogram, Ellipse, and Trapezium. Since Square is treated as a special case of the rectangle, orientation checking is similar to the rectangle. On the other hand, for Circle, orientation checking is not necessary as this shape does not affect the above process of recognizing shapes. For Loop and Line, the proposed method uses an iterative procedure and the projections on the principal axis as presented earlier, respectively. It is noted that these two procedures are invariant to rotation. Sample illustration of the process of recognizing shapes, which involves rotating edge components to zero orientation, estimating mask boundary and finding overlapping region between the component boundary and mask boundary for rotated/tilted edge components are shown in Figure 3.11 (a) and Figure 3.11 (b), respectively.



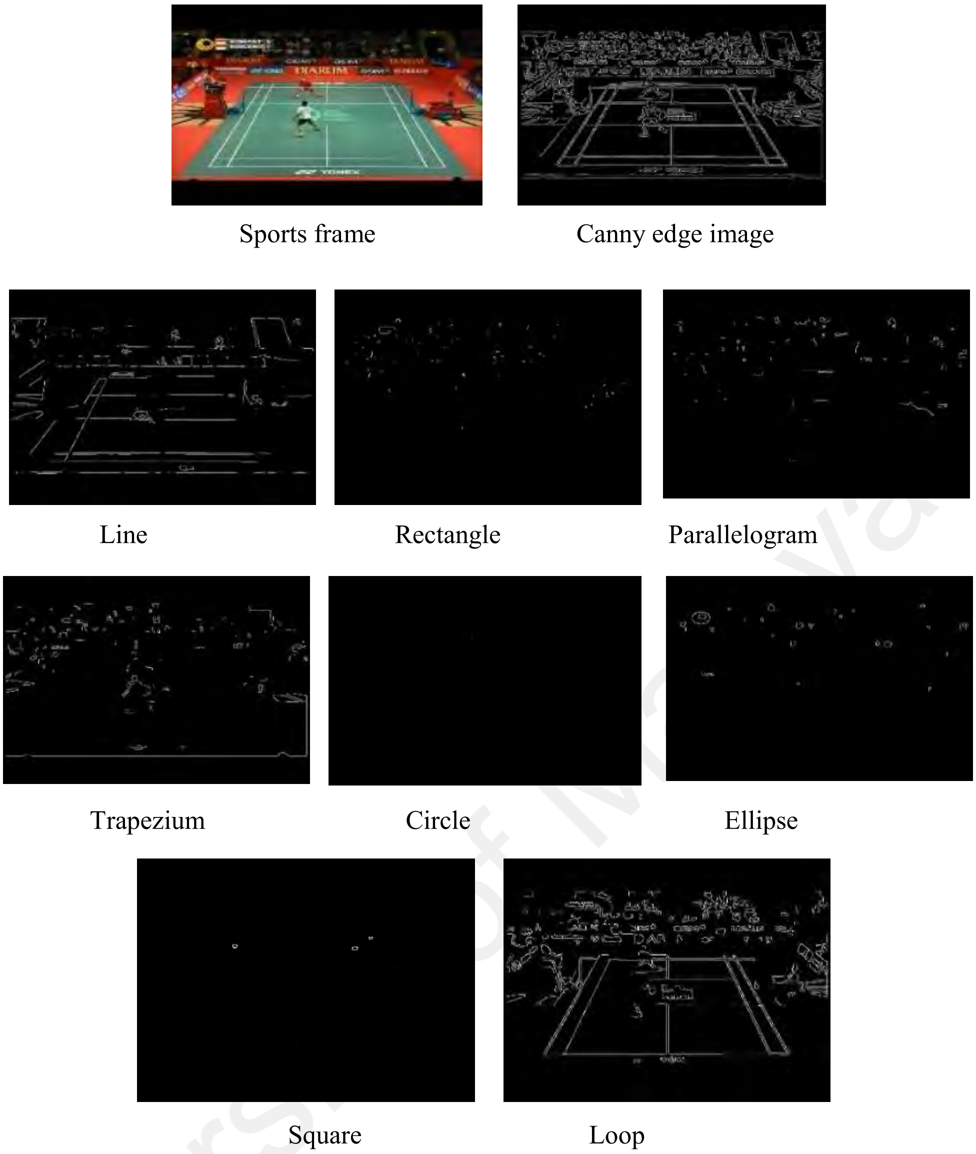
(a) Rectangle (R)

(b) Parallelogram (P)

Figure 3.11: Fixing mask boundary and recognizing shapes for rotated or tilt edge components.

Note: for the purpose of visualization, morphological operation is performed for the edge components. In case of mask boundary results, blue color denotes interior of mask boundary. In case of approximation, green color denote edge component boundary, while the white region represents the overlapping between edge boundary and mask boundary.

Sample edge components for grouping according to the shapes defined above are shown in Figure 3.12, where it can be seen that the edge components are classified according to the definitions of geometrical shapes. As mentioned at the beginning of the section, it is preferred to use Canny edge detector because it has the ability to preserve structures of edge components and to generate fine edges for both low contrast and high contrast frames, which is considered in this work as shown in Figure 3.12. From Figure 3.12, it is observed that edge components for all the groups in spite of low contrast input video frames. To know the effect of the Canny edge detector, Sobel edge detector is compared for the same input frame to classify edge components into respective groups as shown in Figure 3.13, where it can be noticed that a few pixels are missing in the groups compared to the groups of the Canny edge detector. This shows that Sobel edge detector loses sometimes edges for low contrast video frames, which leads to poor performances. Section 3.4 presents experimental evidence to support the statement.



Sports frame

Canny edge image

Line

Rectangle

Parallelogram

Trapezium

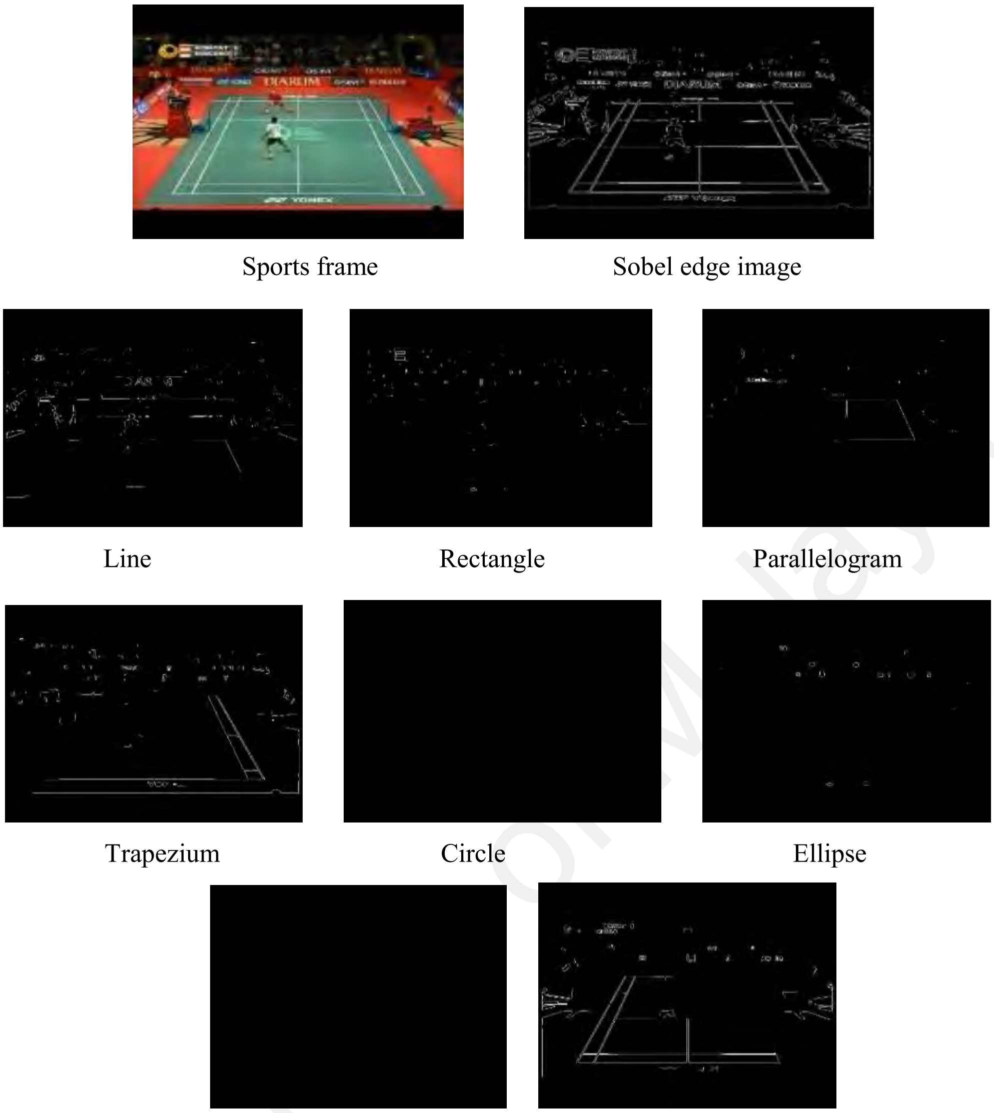
Circle

Ellipse

Square

Loop

Figure 3.12: Sample components grouping based on shape analysis for the Canny edge image of the sports frame using rough-fuzzy.



Sports frame Sobel edge image

Line Rectangle Parallelogram

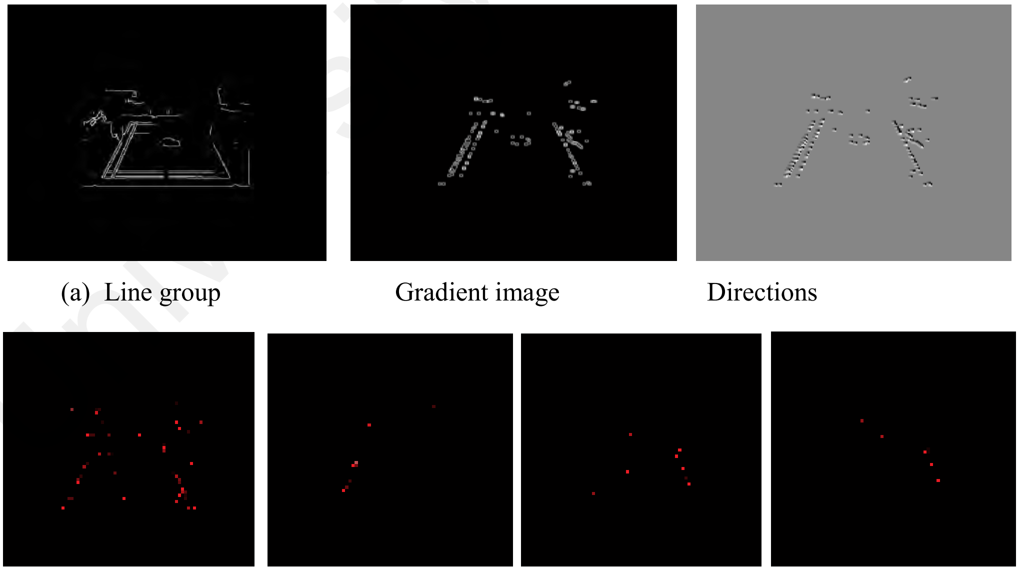
Trapezium Circle Ellipse

Square Loop

Figure 3.13: Sample components grouping based on shape analysis for Sobel edge of the input sports frame with rough-fuzzy.

3.3.2 Gradient-based Intra and Inter Feature Extraction for the Groups

For each group given by the above-presented method in the previous section for video frames of every class, gradient direction and values are explored for extracting distinct features to classify frames. Motivated by the work proposed in (M.-K. Zhou et al., 2016) for recognizing handwriting characters, where it is shown that stroke distribution provides the vital clue for recognizing different handwriting styles of characters, the gradient directions is explored by distributing pixels into a number of planes according to gradient direction to find local distribution of pixels. It is illustrated in Figure 3.14, where (a) shows the sample image of Line group, its gradient image, and gradient directions, and (b) shows pixel distribution into a number of planes according to gradient angle, which generally varies from -180 to + 180. This process results in angular planes for each group. For each angular plane, k-means clustering is used to find different clusters as depicted in Figure 3.15, where clusters with different colors can be seen for the P1 plane in Figure 3.14 (b). This helps in finding the relationship between the pixels in each plane. Here the relationship is defined as how the pixels are close to each other in terms of contrast. Then the proposed method computes the mean, median and standard deviations for each cluster, which gives 3 features for each plane. The relationship among the pixels in each plane, called intra plane, is encoded by covariance and correlation as defined in equation (3.12) and equation (3.13). Inspired by the work in (M.-K. Zhou et al., 2016), the same features are proposed to use for extracting feature vectors of the planes. This process of feature extraction results in a 3600-dimensional feature matrix for each input frame.



(a) Line group Gradient image Directions

P1-Angle: 180 P2-Angle: -162 P7-Angle:- 135 --------P16-Angle:-19

(b) Planes division according to gradient directions.

Figure 3.14: Example of plane generation according to gradient direction to extract structural features (Best viewed in PDF).

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})}{n} \tag{3.12}$$

where X and Y are real valued random variables.

$$R(X,Y) = \frac{cov(X,Y)}{\sqrt{cov(X,X) * cov(Y,Y)}} \tag{3.13}$$

The feature extraction process is formulated as follows. Let the feature matrix (M) be of the size of $p \times 15$, where p is the number of the total angular planes in the frame, while 15 is the dimension of feature vectors (k = 5 clusters and their 3 features). To apply covariance according to equation (3.12), the mean and deviation are calculated for each feature vector. This yields a matrix of size $p \times 15$ using equation (3.14):

$$mv_{p\times 15} = M - II'M(1/p) \tag{3.14}$$

where I is a $p \times 1$ column vector of ones, and I' is the transposed matrix of it. After transposing $mv_{p\times 15}$ and multiplication is done with the original matrix M using the following equation (3.15):

$$MV_{15\times 15} = mv'_{p\times 15} \times mv_{p\times 15} \tag{3.15}$$

It generates a feature matrix of size15×15. And finally, the covariance matrix is obtained by dividing with the number of planes p as defined in equation (3.16):

$$cov\_M_{15\times 15} = MV_{15\times 15}/p \tag{3.16}$$

In the same way, the correlation feature matrix of size 15×15 is reckoned using matrix multiplication and division according to equation (3.13). As a result, for the angular plane P1 in Figure 3.14 (b), the proposed method gives 5 clusters as shown in Figure 3.15. Therefore, the proposed method obtains 225 features using covariance and 225 features using correlation for each group. For 8 groups given by Section 3.3.1 of the frame in Figure 3.7 (a), the proposed method obtains 3,600 features, called as intra plane features since it uses each plane separately. Here the value of *k* is 5 for intra plane features.

In order to find the value for *k* automatically, the following procedure is proposed. 100 samples from each of the 10 classes are chosen randomly, which gives 1,000 frames for determining the value of *k* automatically. For each frame, the proposed method obtains a number of planes as discussed above for the Canny image. For each plane, k-means clustering is applied with *k*=*n*, which gives different numbers of clusters (the value of *k*) for each plane as shown in Figure 3.16 (a) for frame 1 (F1) and frame 2 (F2), where different *k* values for different planes are seen. To find the common cluster number which represents the planes of the frame, histogram operation is performed for *k* values of planes as shown in Figure 3.16 (b), which is called a local histogram. The proposed method chooses the value which gives the highest peak as *k* value at the plane level that should be represented at the frame level. In order to find *k* value at frame level, which represents the whole database, the same histogram operation is performed on *k* values chosen for each sample frame to choose the value that contributes to the highest peak as shown in Figure 3.16 (c) as *k* value for intra plane features, where *k*=5 gives the highest peak.

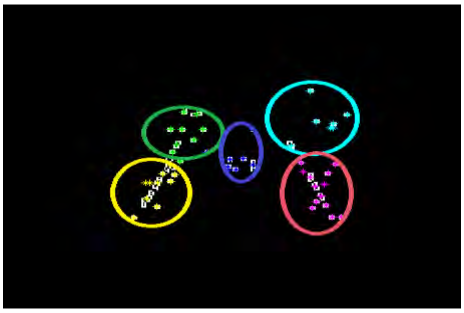
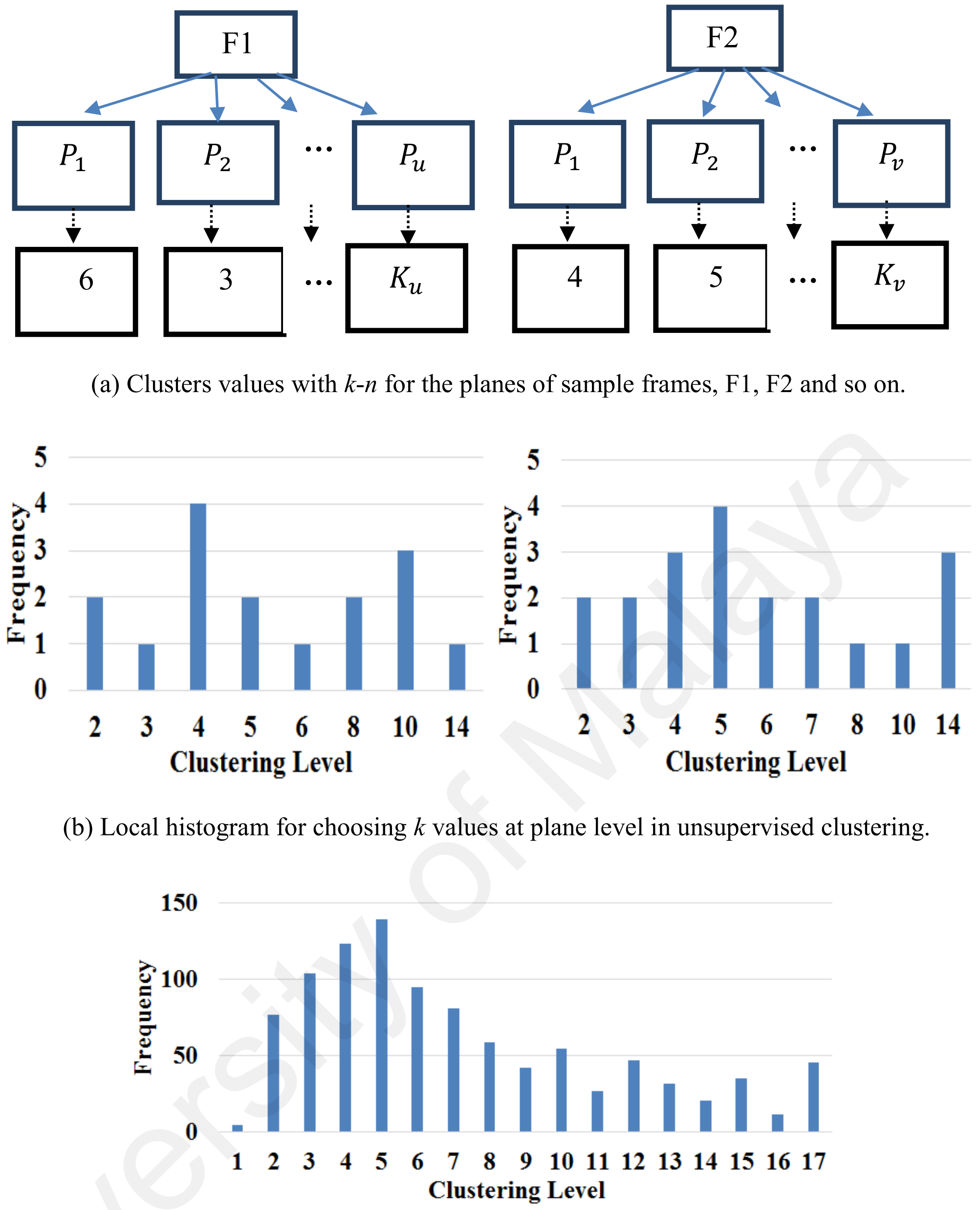


Figure 3.15: Different clusters for plane P1 in Figure 3.14 (b).



(a) Clusters values with *k-n* for the planes of sample frames, F1, F2 and so on.

(b) Local histogram for choosing *k* values at plane level in unsupervised clustering.

(c) Global histogram for choosing *k* values at frame leve

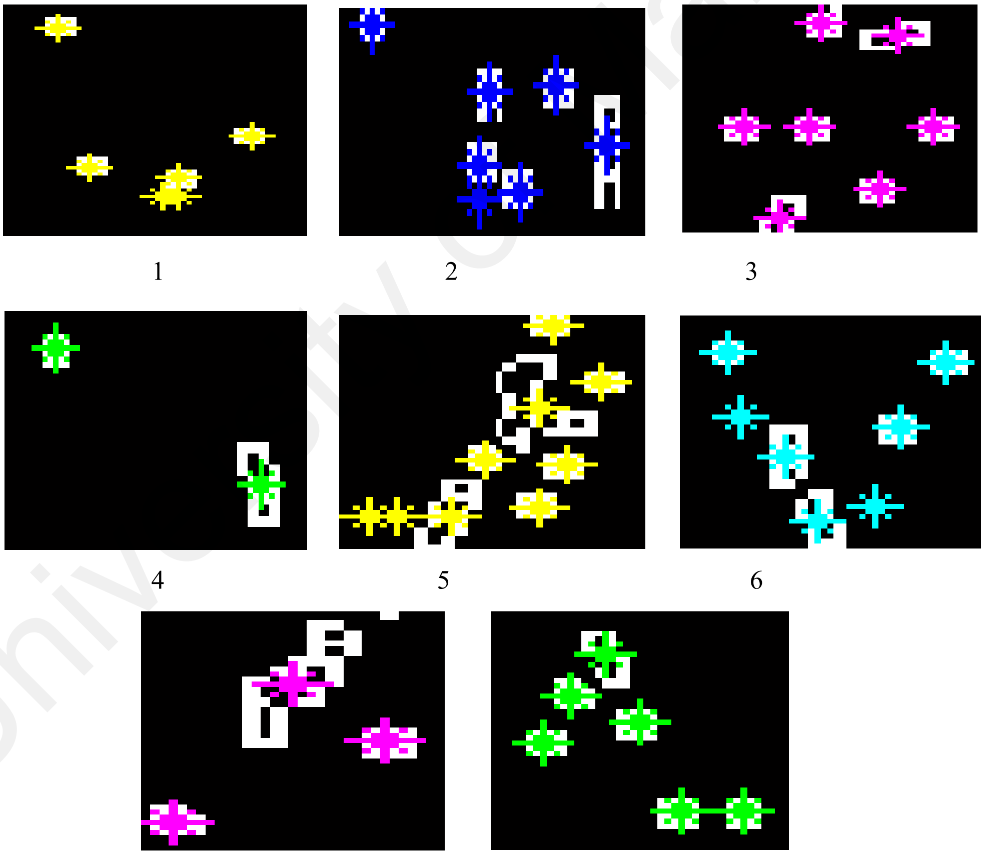
Figure 3.16: Determination of the value for k automatically for intra plane features.

Intra plane features are extracted in the previous section to find the relationship between plane pixels. This leads to extract features across planes to strengthen feature extraction to solve the complex scene type video categorization problem. The number of clusters for planes is determined with the predefined samples. For each plane, k-means clustering has been applied to obtain clusters as shown in Figure 3.17, where 8 clusters for plane P1 can be seen. For the second plane also, the proposed method obtains clusters.

To find the relationship between inter planes, gradient values are extracted across the planes corresponding to the edge components in the clusters. For those gradient values, a histogram is plotted to find the value which contributes for the highest peak as the feature of the particular cluster. This gives a feature vector for one cluster. In the same way, the proposed method obtains feature vectors for other clusters, which forms a feature matrix. The covariance and correlation are estimated for the feature matrix as mentioned in the previous section, which gives 64 features (8×8) for each group given by the method presented in Section 3.3.1. In total, since *k* is 8, the number of features would be 64×8 = 1024 for each frame. As discussed in the previous section, to determine the value for *k* automatically, the following procedure is proposed using the same 1, 000 samples. The proposed method obtains Canny edge image (C) for the samples as shown in Figure 3.7. Figure 3.18 (a), then it applies k-means clustering with *k*=*m* for each edge image as shown in Figure 3.18 (a), where different *k* values for different frames are observed. To choose the value for *k*, the histogram is performed on *k* values as shown in Figure 3.18 (b), where it is noticed 8 is contributing to the highest peak and hence it is considered as the actual value of *k*. For the input frame, 3600+1024 = 4624 features are extracted for classification.

In summary, algorithmic representation for inter-plane feature extraction and determining the value of *k* are presented below. The steps in training phase describe how to determine the number of clusters *k*, which is the parameter of k-means clustering using predefined samples. For each sample, the proposed method obtains Canny edge images. Then the same clustering technique is employed on all the Canny images, which outputs a number of clusters (the value of *k*) for each sample. To choose *k* value which represents the whole database, histogram analysis is performed on *k* values obtained for each sample. The value which contributes to the highest peak is considered as the actual *k* value of k-means clustering at the frame level. Similarly, the steps in testing phase describe how to extract features for testing samples. For each testing sample, the proposed method obtains

8 groups using rough-fuzzy combination method presented in Section 3.3.1. Each group is divided into angular planes based on gradient direction of components in the group. For each angular plane, the proposed method employs k-means clustering with k (determined earlier) value on each plane, which results in k clusters. The proposed method extracts gradient values across the planes corresponding to edge components in the clusters. For those gradient values, a histogram is plotted to find the value which contributes the highest peak as the feature vector of the particular cluster. The covariance and correlation features are extracted for the feature vectors, and this results in a feature matrix for classification.

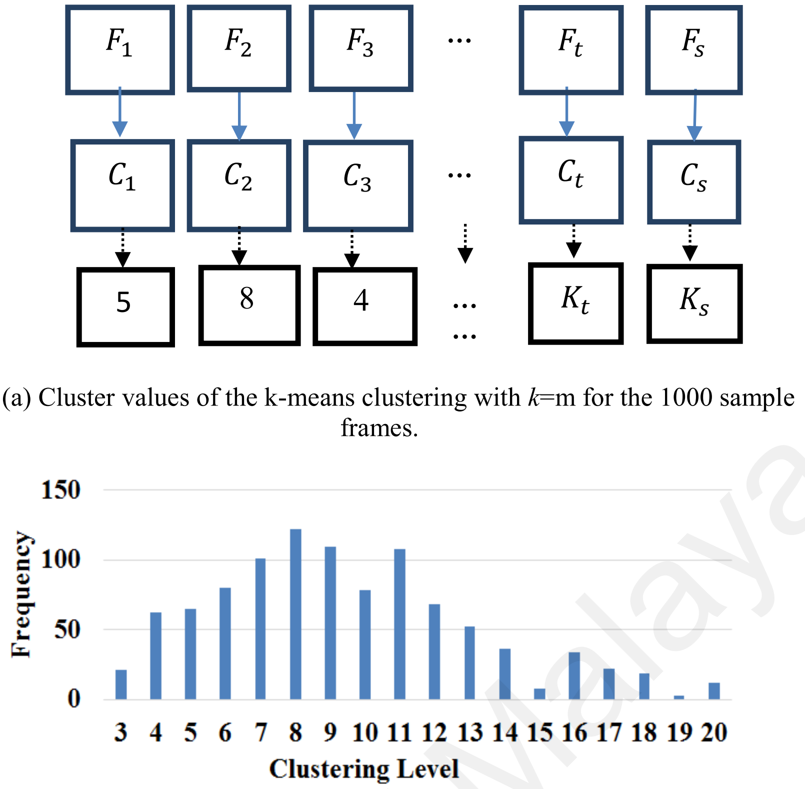


1 2 3

4 5 6

7 8

Figure 3.17: Inter-plane feature extraction for Plane P_1.



(a) Cluster values of the k-means clustering with *k*=m for the 1000 sample frames.

(b) Histogram for choosing k value automatically.

Figure 3.18: Determination of k value for inter plane feature extraction.

Algorithm 3.1: Training for video classification

```
Input: Gray color image of the input image

1. For each  gray image in training database:
   a. Apply canny edge operator on the gray image.
   b. Apply unsupervised clustering on the canny image to find clustering number.
2. Find the clustering number ($k_{Inter}$) having more frequency using $\max_{c}\left(\sum_{i=1}^{C} h_i\right)$
   from all the images,
   Where $i$ represents the unique cluster obtained from Step 1, while $h_i$ denotes the
   total contributed clusters in the $i_{th}$ cluster.

Output: Cluster number ($k_{inter}$)
```

Algorithm 3.2: Testing for video classification

Input: Testing images in gray color and k_{Inter} obtained by Algorithm 3.1, where unsupervised clustering has been applied to find the number of clusters.

1. Apply Rough Fuzzy component grouping method on gray image to obtain 8 groups using the steps presented in Section 3.3.1 and shown in Figure 3.12.
2. For each component group image shown in Figure 3.12,
    a. For every pixel, calculate $\theta = \tan^{-1}\frac{y}{x}$ to find angular planes, where all the pixels having the same angle belong to one angular plane.
    b. For each unique angular plane, apply k-means clustering, where the number of clusters has been set to k_{inter} obtain from the training (Algorithm 3.1).
    c. For each clustering group, across all the angular planes, and estimate the maximum frequency (v) of gradient component using $\max\limits_{\mu_l,\theta}\left(\sum \tan^{-1}\frac{Y}{X}\right)$ described in Section 3.3.2 (see Figure 3.14).
    d. Compute covariance matrix (COV) using $\frac{1}{k_{Inter}}\sum_{l=1}^{k_{Inter}} v\ (v)^T$
    e. Compute correlation matrix (COR) using $\frac{\mathrm{Cov}_i}{\sqrt{\frac{1}{k_{Inter}}\sum_{l=1}^{k_{Inter}} v\ (v\ )^T}}$

Output: Covariance (COV) and correlation(COR) matrix.

3.3.3 Feature Extraction using Temporal Frames for Video Categorization

Since the input video provides temporal frames, temporal information is exploited to increase the discriminative power of the feature extraction in this work. For the extracted features as discussed in the previous section, the same features are extracted from the left and right sides of the key frame if available. Otherwise, the proposed method considers three consecutive frames for feature extraction. It computes the average of the three feature matrices given by three frames, which gives the final feature matrix for classification.

It is noted from the literature that Neural Network (NN) is a nonlinear model and has the ability to identify complex nonlinear relationships between dependent and independent variables. As a result, it is a non-parametric model compared to parametric models that require higher statistical calculation. Although there are two other types of neural networks, namely, Radial Basis Function (RBF) networks and Learning Vector Quantization (LVQ) networks, feed forward perceptron trained with backpropagation is used in solving problems for its higher degree of generalization from training data. It is noted that the feedforward neural network classifier used in (Huiping Li, Doermann, & Kia, 2000) explores the above characteristics of NN for classifying text and non-text pixels in videos, thus the NN classifier is proposed in the same way for classification in this work (Huiping Li et al., 2000).

Since the problem is 10 class classification, 10 output nodes are considered, one for each of the ten classes. Two intermediate layers are used in this classification. Every node on one layer is connected with the nodes on the previous layer. The output of a node is defined as a function of the weighted sum of the connected nodes in the previous layer. Here, neural network considers random values as the initial weights and then updates the weights automatically during learning stage according to problems. For choosing training samples, a 10-fold cross validation procedure is used, which automatically provides the number of training and testing samples for classification. In this work, 30,000 frames are considered for classification, including the temporal frames. Out of which, 27,000 frames are used for training.

3.4 Experimental Results and Comparative Study

In this section, the datasets used for two approaches are described. The evaluation metrics have been explained in terms of image/video classification experiments, text detection experiments through F-score, and recognition experiment through binarization. Comparative studies are performed for validating the efficiency of the presented and existing classification. In addition to these, comparison of detection and recognition results has also been tabulated for “before classification” and after classification”.

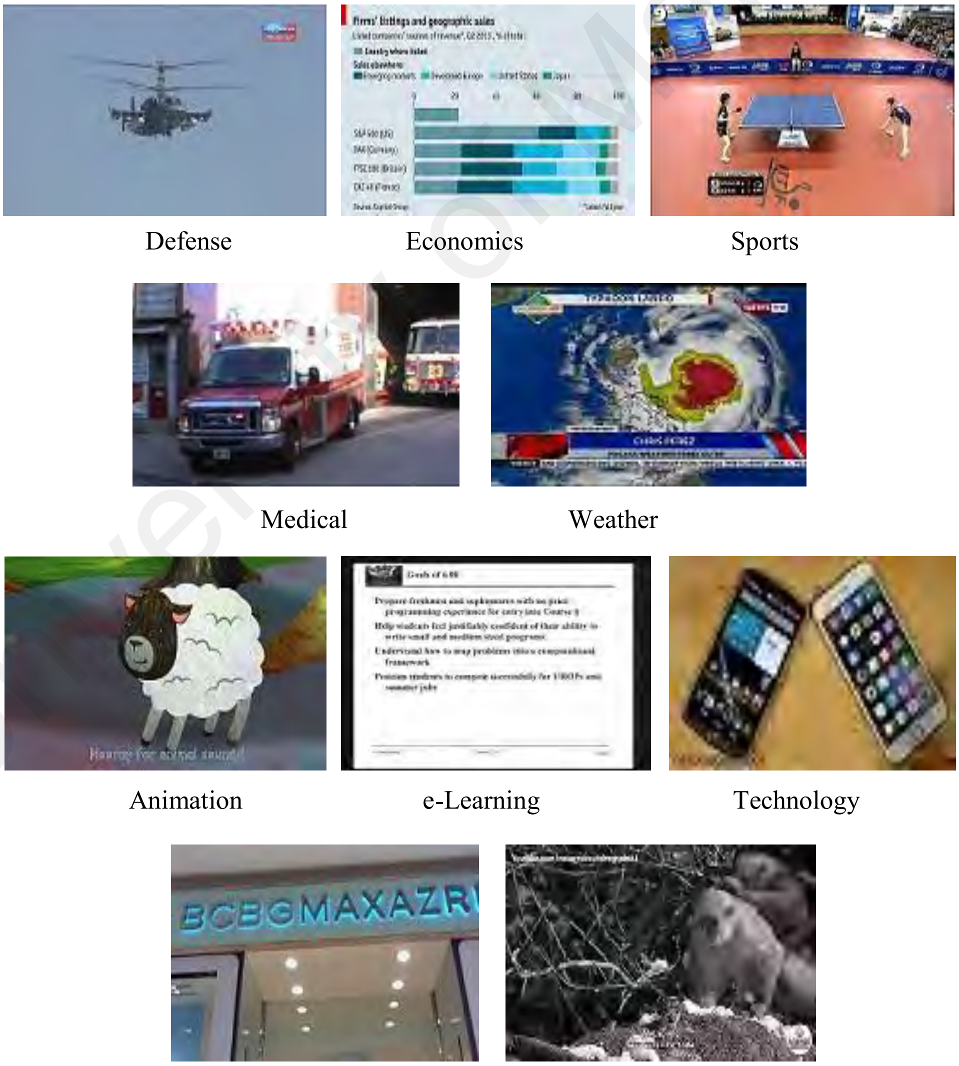
Section 3.4 is organized as follows. Section 3.4.1 discusses the dataset and evaluation metrics, Section 3.4.2 and Section 3.4.3 illustrate the results of the proposed approach and finally section 3.4.4 gives the details comparison of proposed and existing classification methods.

3.4.1 Datasets and Evaluation

Video data is collected from YouTube multi-oriented scene texts with complex background images as there are no standard datasets available for the categorization of different scene type video in literature. The video classes are chosen according to the important role in the smart city and digital city development (Rathore, Ahmad, Paul, & Rho, 2016). The database includes the frames with the resolution ranging from 480×360 to 1920×1080 pixels, caption texts, and scene texts with variant fonts, orientations, backgrounds, contrasts, etc. The dataset used for Fuzzy-mass based approach is different than the dataset used for Fuzzy-rough in respect to size and class number. Furthermore, later uses temporal frames whereas former takes a single frame for classification.

For Fuzzy-mass based approach, namely, Defense, Economics, Sports, Medical, and Weather are considered. For this approach standard databases such as NUS data (N. Sharma, Shivakumara, et al., 2012), and YVT video data (Nguyen, Wang, & Belongie, 2014) are considered as there are some video frames similar to the chosen class. In summary, 120 for Defense, 100 for Economics, 108 for Sports, 75 for Medical and 100 for Weather, are collected from different sources of video. This gives total 503 frames for experimentation. The sample frames of five classes, namely, Defense, Economics, Sports, Medical, and Weather, are displayed in Figure 3.19, where different backgrounds and foregrounds are noticed for different classes.

For Fuzzy-Rough based approach, more 5 classes are included with existing classes. YouTube and other internet sources are used for collecting the dataset for 10 classes with temporal frames. The new classes are namely Animation (A), e-learning (e-L), Technology (T), Outlet (O), and Animal Planet (AP), to evaluate the proposed classification method as each class consists of 3,000 frames, which includes three temporal frames for each keyframe. For the 10 classes, 30,000 frames are obtained for experimentation in this work. The considered large data are close to generalized data for the above mentioned 10 classes. As discussed in the beginning of the chapter, each video classes poses different challenges, like low resolution, contrast, font, font size and background variations, multi-oriented texts, etc., due to different nature and characteristics.



Defense Economics Sports

Medical Weather

Animation e-Learning Technology

Outlet Animal

Figure 3.19: Samples of dataset and successful classification results of the proposed method.

The presented approach involves three types of experiments to evaluate the performance of the proposed method: classification experiments, text detection experiments, and recognition experiment through binarization. For evaluating the proposed classification step, standard classification is calculated rate through confusion matrices. For evaluating text detection results of different text detection methods, the instructions are followed given in (Epshtein et al., 2010; Khare, Shivakumara, & Raveendran, 2015; Huiping Li et al., 2000; Liang et al., 2015a; Mosleh et al., 2013; Rong et al., 2014; Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010; L. Wu et al., 2015; G. Zhao & Pietikainen, 2007), which use standard metrices, namely, Recall (R), Precision (P), and F-measure (F). For recognition experiments, Recognition Rate (RR) is calculated at character level for different binarization methods. Experiments, namely, prior to classification and after classification are conducted to verify that the text detection and recognition methods generate low scores for prior to classification and significant improvement after classification. Prior to classification considers frames of all the 5classes (for Fuzzy-Mass concept) or 10 classes (for the Fuzzy-Rough concept) as the input for experiments, while after classification considers individual classes as the input for experimentation. For experiments prior to classification, default parameters are used in the text detection methods to calculate R, P, and F. However, for the experiments after classification, since classes are known by the classification methods as discussed in Section 3.3.1, the parameters are tuned based on training samples of each class to calculate recall, precision, and F-measures. In general, after classification, text detection and recognition performance improve significantly because by considering the advantage of classification.

In the same way of text detection experiment for each classification methods, average Recognition Rate (RR) is reported of the different binarization methods for the classes of respective classification methods. Key parameters are determined from each binarization method listed in Table 3.21 and Table 3.23, such as threshold value for the Bayesian classifier (S. Roy et al., 2015), the threshold value for Canny edge image (Howe, 2013), and window sizes for different binarization algorithms (Milyaev et al., 2013). Since Su et al. (Su et al., 2013) provided exe files that find values automatically, there is no option for tuning.

3.4.2 Experiments on Fuzzy-Mass based Approach

The classification rates of local, global and combined (local+global) experiments are reported in Table 3.1, Table 3.2 and Table 3.3, respectively. According to the experiments reported in and Table 3.2, it is observed that local features give better results (the average accuracy is 70.52%) than global features (the average accuracy is 48.3%). This is true because local features have a more discriminative power than global features due to the non-uniform distribution of edges in the edge image of a frame. When Table 3.1, Table 3.2 and Table 3.3 (for proposed method) are compared, the combined (local+global) gives better results (the average classification rate is 78.3%) than local alone or the global alone because the combined method utilizes the advantages of both local and global features.

Table 3.1: Confusion matrix using only Local features.

class	D	E	S	M	W
D	78.3	2.0	14.7	0.5	4.2
E	9.6	89.7	0.1	0.1	0.4
S	8.4	0.4	85.3	0	5.7
M	29.4	6.8	20.9	15.1	27.6
W	2.7	0	11.0	1.9	84.2

Table 3.2: Confusion matrix using only Global features.

class	D	E	S	M	W
D	36.9	22.5	29.5	2.0	8.9
E	9.0	81.2	7.6	0.2	1.7
S	1.8	7.0	46.6	2.3	42.0
M	2.2	9.3	22.9	8.8	56.5
W	0.8	4.5	23.1	3.4	67.9

Table 3.3: Confusion matrix using Local +Global features.

Class	D	E	S	M	W
D	75.4	1.5	22.4	0.3	0.2
E	0.3	99.6	0	0	0
S	0.1	0	96.2	0	3.6
M	3.3	2.3	20.2	62.2	11.7
W	0	0.4	4.7	3.1	91.7

3.4.3 Experiments on Rough-Fuzzy based Approach

It is noted that the proposed method involves the following key steps for the classification of scene type videos, namely, covariance and correlation feature extraction for classified edge components by the combination of rough set and fuzzy, the use of intra and inter planes, and the use of Sobel edge images and Canny edge images. In order to analyze the contributions of the above key steps, experiments are conducted as follows, (1) covariance+intra+inter+classification, (2) correlation+intra+inter+classification, (3) covariance+correlation+intra+classification, (4) covariance+correlation+inter+classification, (5) covariance+correlation+intra+inter+classification using Sobel edges of the input frames, and (6) covariance+correlation+intra+inter+classification using Canny edge of the input, which are required to analyze the contributions of covariance, correlation, intra plane features, inter plane features, Sobel edge detector and Canny edge detector, respectively. For each experiment, confusion matrices are respectively estimated as reported in Table 3.4 - for the above 6 experiments. The average classification rate, which is the mean of diagonal elements of the confusion matrices for the respective 6 experiments are 55.6%, 58.9%, 54.0%, 63.0%, 56.3% and 76.0%. This shows that all the 6 key steps contribute almost equally except the features using inter planes and the proposed method with Canny edge detector. This indicates the features extracted from the edge components corresponding to clusters across planes have more discriminative power than intra planes. However, the introduced approach with Canny and temporal information achieves the best average classification rate (76.0%) compared to the same approach with Sobel and temporal information (56.3%). It is true that Sobel edge detector is good for high contrast images as it involves the first order derivative with one optimal threshold, while Canny is good for high and low contrast images as it involves double optimal thresholds. At the same time, the considered dataset contains images of different contrasts variations. Therefore, the accuracy with the Canny edge is better compared to Sobel edge. There is a significant difference when the average classification rates of covariance, correlation, intra and inter are compared with the proposed method (76.0%). This is due to the integration of strengths of covariance-correlation among pixels in intra and inter angular planes given by rough-fuzzy combination, which extracts unique stroke distributions locally and globally from irregular edge patterns in edge components of frames, and temporal information which adds stability to features by considering information in neighboring frames. Therefore, the proposed idea aggregates the advantages of a new way of combination of rough-fuzzy for grouping edge components, covariance-correlation of intra, inter planes, Canny edge detector, and temporal information of video to accomplish better accuracy for the complex classification. Samples of successful classification are portrayed in Figure 3.19.

Table 3.4: Confusion matrix of covariance+intra+inter+classification.

Class	D	Ec	S	M	W	A	E-l	T	En	AP
D	48.1	9	3	11.3	4.1	2.2	3.2	4.1	4.7	10.3
Ec	4.6	56.7	4.2	3.5	7.3	2.6	6.1	6.3	3.6	5.1
S	10.4	1.1	50.6	4.1	7.2	4.9	1.8	13.3	3.9	2.7
M	2.1	3.7	4.2	52.2	1.1	15.2	0.7	2.3	5.6	12.9
W	4.1	2.9	1.4	3.8	55.8	18.8	7.7	2.9	0	2.6
A	11.9	3.7	8.9	4.8	1.2	50.1	5.1	0	0	14.3
E-l	2.1	0	3.7	6.2	12.8	3.5	57.8	5.2	5.1	3.6
T	1.4	19.2	2.6	4.8	3.9	2.8	0.8	54.9	6.1	3.5
En	16.2	3.8	2.4	1.3	0	2.1	4.1	4.8	65.3	0
AP	3.6	5.8	2.8	4.9	5.3	4.3	2.7	2.4	4.1	64.1

Table 3.5: Confusion matrix of correlation+intra+inter+classification.

Class	D	Ec	S	M	W	A	e-L	T	O	AP
D	40.3	1.2	11.4	2.8	1.9	4.1	4.8	13.3	3.9	16.3
Ec	9.1	63.8	0.9	1.2	2.3	6.7	8.1	1.2	4.6	2.1
S	4.2	2.8	66.2	1.1	3.1	1.4	10.8	5.9	2.1	2.4
M	7.3	1.6	3.3	69.2	3.7	4.1	5.2	2.3	2	1.3
W	8.4	3.9	7.1	1.8	61.4	4	1.3	4.7	7.4	1
A	5.8	1.8	3.4	11.3	1.2	56.7	4.9	0.4	1.9	12.6
e-L	5.2	2.1	19.1	0.2	1.1	1.4	54.7	12.4	2.8	1
T	0	12.3	1.5	10.4	2.8	15.3	2.1	50.6	3.3	1.7
O	2.5	4.1	7.1	3.1	1.9	3.9	6.1	5.5	64.5	1.3
AP	11.7	1.2	0.6	1.3	0	2.7	3.1	5.7	11.9	61.8

Table 3.6: Confusion matrix of covariance+correlation+intra+classification.

Class	D	Ec	S	M	W	A	E-l	T	En	AP
D	49.8	3.4	1.3	2	2.5	3.2	13.4	8.3	5.9	10.2
Ec	3.5	42.1	3.8	2.9	14.8	10.3	3.9	12.1	2.8	3.8
S	7.1	2.8	56.7	4.5	3.9	4.6	3.9	9.8	3.1	3.6
M	3.6	3.1	8.2	57.1	3.4	2.8	9.8	4.8	2.3	4.9
W	2.8	1.3	3.8	12.9	59.1	8.5	2.4	1.7	1.7	5.8
A	6.2	1.8	2.8	2.4	4.5	51.2	6.9	1.4	10.2	12.6
E-l	14.9	4.8	2.5	0	1.3	1.7	57.8	2.8	2.8	11.4
T	10.8	6.8	0	1.2	3.2	2.7	2.9	49.8	19.7	2.9
En	0.5	12.5	1.8	4.6	1.2	11.9	3.1	1.4	56.2	6.8
AP	10.9	1.8	2.1	2.5	0.7	3.1	3	3.5	11.3	61.1

Table 3.7: Confusion matrix of covariance+correlation+inter+classification.

Class	D	Ec	S	M	W	A	e-L	T	O	AP
D	60.1	10.3	1.4	2.1	12.6	2.6	1.8	2.1	2.9	4.1
Ec	4.7	66.7	4.7	2.9	1.2	3.3	12	0.2	2.1	2.2
S	0	2.9	69.6	1.3	8.4	1.1	6.2	0.2	7.1	3.2
M	3.7	2.8	3.9	65.5	2.1	2.8	1.9	0.4	10.8	6.1
W	3.1	4.9	4.3	1.6	59.8	0.7	0.4	12.8	5.2	7.2
A	10.1	1.9	3.5	1.7	0.6	62.2	2.8	3.9	10.8	2.5
e-L	3.7	2.8	2.7	0.8	1.8	14.3	56.3	7.2	0.4	10
T	10	4.4	1.8	1.3	2.1	1.7	3.1	64.6	8.9	2.1
O	7.7	3.9	1.7	2.9	9.1	1.1	0.5	3.1	67.4	2.6
AP	13.1	2.6	3.1	13.3	0.5	2.1	3.1	2.4	1.2	58.6

Table 3.8: Confusion matrix of the proposed method using Sobel edge components with temporal frames.

Class	D	Ec	S	M	W	A	e-L	T	O	AP
D	61.2	2.3	1.2	3.8	11.2	6.3	5.1	4.6	3.1	1.0
Ec	5.3	58.7	1.3	2.3	1.9	4.8	2.9	1.0	7.3	14.2
S	8.0	1.0	52.9	16.0	5.0	2.1	3.3	4.8	2.5	4.0
M	2.5	3.0	5.0	67.0	2.4	3.9	0.3	1.9	11.8	2.9
W	6.3	2.0	8.2	3.0	49.1	2.7	4.0	12.7	6.0	5.5
A	2.1	10.0	15.7	1.2	12.9	44.2	5.3	3.7	2.0	2.5
e-L	5.2	2.4	6.0	4.3	3.0	7.5	55.0	12.0	2.3	1.9
T	11.0	1.2	2.0	7.6	13.0	2.6	5.2	50.0	1.8	5.2
O	4.1	3.2	2.4	6.8	8.0	4.2	1.0	6.0	63.3	0.7
AP	2.1	1.3	5.2	13.0	2.4	3.2	1.9	1.0	7.5	62.0

Table 3.9: Confusion matrix of the proposed method using Canny edge components with temporal frames.

Class	D	Ec	S	M	W	A	e-L	T	O	AP
D	83.9	1.03	2.7	1.7	2.23	2.3	1.07	1.6	2.4	1.04
Ec	1.3	75.5	2.04	1.04	1.7	2.67	4.22	1.53	5.8	4.2
S	2.21	6.47	72.7	2.8	4.7	1.05	3.9	2.4	1.05	2.69
M	5.8	1.36	2.8	71.5	4.06	6.53	1.23	2.12	3.07	1.49
W	1.8	2.3	1.32	1.17	80.6	3.5	1.47	2.53	3.71	1.56
A	1.4	3.26	2.51	2.4	1.1	77.7	2.78	4.9	2.41	1.54
e-L	1.4	3.26	2.51	2.4	1.1	1.78	78.7	4.9	2.41	1.54
T	1.83	3.28	6.9	1.07	1.82	1.32	2	75.8	1.78	4.2
O	12.9	1.82	2.96	1.06	1.93	2.45	1.48	3.73	69.5	2.17
AP	1.37	2.78	1.05	2.67	6.58	4.72	2.47	6.48	2.38	69.5

3.4.4 Comparative Study and Discussion

To show the efficacy of the presented classification, the recent video classification methods are implemented as per the instructions given in these papers and use the same experimental set up as the proposed method for comparative studies, namely, Bosch et al.'s method (Bosch et al., 2008) which explores probabilistic latent semantic analysis and SIFT features for scene image classification, Dunlop's method (Dunlop, 2010) which proposes scene classification of images and videos through semantic segmentation, and Qin et al.'s method (Qin et al., 2016) which proposes statistical, structural and spatial features in color space with an SVM classifier for video text frame classification. The reason to choose the above three methods for comparative studies is that Bosch et al.'s method focuses on scene image classification, Dunlop's method focuses on video classification, which utilizes temporal frames like the proposed method, and Qin et al.'s method focuses on video text frames as the proposed method.

To show the advantage of classification, three types of text detection methods are used/implemented: the methods that exploit temporal frames, namely, Khare et al. (Khare, Shivakumara, & Raveendran, 2015), Moselh et al. (Mosleh et al., 2013), Zhao et al. (X. Zhao et al., 2011) and Li et al. (Huiping Li et al., 2000), the methods that do not use temporal information, namely, Shivakumara et al. (Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010; Shivakumara et al., 2012), and the methods that are developed for scene images, namely, Yin et al. (Yin et al., 2014), Rong et al. (Rong et al., 2014) and Epshtein et al. (Epshtein et al., 2010). The text detection methods, developed for scene images are considered for experimentation because the video contains scene texts in the complex background, which share the same characteristics as in natural images. Similarly, three types of binarization methods, namely, Roy et al. (S. Roy et al., 2015), developed for binarizing text in natural scene images, Milyaev et al. (Milyaev et al., 2013), developed for binarizing text in the degraded document images, Su et al. (Su

& Lu, 2014) and Howe (Howe, 2013) have been used. The main reason to choose the three types of methods is that the considered scene type video classification is a complex problem which suffers from different challenges, such as contrast and background variations. As a result, the challenges influence directly on text in frames.

To validate the above analysis, the proposed method is compared using GOOGLE API (Cloud Vision, 2011), which is available publicly and uses deep learning, cloud, and a large number of features for retrieving scene images that contain multiple objects in each image. The purpose of doing experiments with this system is to show that the performance of the systems which involve deep learning that depends heavily on a number of labeled samples and setting optimal parameters to achieve good results. Besides, this set up may not be feasible for the data which consists of a small number of samples, or when background complexity varies greatly for samples of the same class. Confidence score is generated for training samples using GOOGLE API. This process gives different labels for each class with confidence scores. For instance, Animal Planet (AP) class can have agriculture, black bird, branch, nature labels, etc., while Animation can have amusement park, amusement ride, atmosphere, etc. Labels are given by GOOGLE API system. In this way, feature vectors are created for all the classes based on training samples. A cut off threshold 85% is set to confidence score to generate the final confusion matrix for all the classes. This 85% cut of is fixed based on the experiments on training samples. It is observed from experiments that if the cut off value is increased, the method includes irrelevant labels and if it is decreased, it loses relevant labels.

The quantitative score of discussed approach is tabulated in Table 3.3, and the quantitative result of GOOGLE API is given in Table 3.13 and Table 3.14. According to analysis, it is noted that the GOOGLE API system works when it recognizes multiple objects correctly in images. If the image contains an object which is not trained by the system, GOOGLE API fails to classify the image correctly, while the proposed method is not trained on specific shapes of objects, rather it studies the pattern of edge components using Fuzzy and rough set combination; thus it gives better results for the dataset. However, if GOOGLE API is trained with the dataset, it may score better results than the proposed method. But this is the limitation of the GOOGLE API system as its performance depends on the number of labeled samples. Moreover, the discussed approach does not require such large number of samples for achieving good results. In addition, according to the website (Engine, 2008) and experiments, it is noticed that GOOGLE API works based on shapes of multiple objects in scene images. However, in case of the collected dataset, one cannot expect particular shapes of objects because scene type images of dataset may contain objects or may not. For example, Weather and Economic scene classes do not contain any object with particular shapes. Therefore, GOOGLE API scores poor results compared to the proposed method.

In case of Fuzzy-Mass based approach, Table 3.10, Table 3.11, Table 3.12 and Table 3.13 include confusion matrices of the four existing methods. It is observed that the presented approach exceeds the existing classification's accuracy. The reason is as follows: Qin et al.'s method (Qin et al., 2016) cannot overcome ambiguity in classification as reported in Table 3.10. The methods in (Cloud Vision, 2011; J. Liu et al., 2016; Z. Xu et al., 2016) require multiple objects with clear shapes for annotating images correctly. However, for the images considered in this work, one cannot expect clear objects, especially for weather, economics and sports ones. Furthermore, the presented method does not depend on object shapes rather it extracts edge patterns of image content in a new way for classification.

For Fuzzy-Rough based approach, the quantitative results of Bosch et al.'s method (Bosch et al., 2008), Dunlop's method (Dunlop, 2010) and Qin et al.'s method (Qin et al., 2016) are reported in Table 3.15 - Table 3.17, respectively. It is observed from Table 3.15 - Table 3.17 that the accuracy of the proposed method is superior to the existing methods. The reason for the poor results of the existing methods is that they require multiple objects to train classifiers. Moreover, the Fuzzy-Rough approach extracts unique shapes from irregular edge patterns by exploring the rough-fuzzy and distinct relationship among pixels locally and globally based on covariance-correlation of intra, inter planes and temporal information.

Table 3.10: Confusion matrices of the existing classification method (Qin et al., 2016).

Class	D	Ec	S	M	W
D	61.1	10.3	8.8	15	4.8
Ec	4.3	90.7	1.3	2.4	1.3
S	3.7	4.8	77.6	4.6	9.3
M	1.03	3.5	5.4	82.1	7.8
W	1.9	1.7	2.8	2.2	91.3

Table 3.11: Confusion matrices of the existing classification method (Bosch et al., 2008).

Class	D	Ec	S	M	W
D	69.2	8.7	5.8	3.4	12.9
Ec	0.8	86.7	8.2	1.2	3.0
S	5	1.57	90.2	2.0	1.2
M	11.03	0.3	2.2	81.2	5.2
W	3.68	2.5	6.2	0.31	87.2

Table 3.12: Confusion matrices of the existing classification method (Dunlop, 2010).

Class	D	Ec	S	M	W
D	75.2	2.7	3.5	4.8	13.7
Ec	0.7	85.6	1.2	5.9	6.5
S	3.7	0.3	94.0	1.5	0.4
M	7.3	1.4	4.2	80.4	6.6
W	7.3	3.9	6.9	2.8	78.9

Table 3.13: Confusion matrices of the existing classification methods (Cloud Vision, 2011) system.

class	D	Ec	S	M	W
D	85.3	3.04	6.2	1.2	4.1
Ec	13.4	63.0	8.03	8.1	7.2
S	7.1	2.8	81.4	3.1	5.6
M	4.2	5.6	0.1	74.0	15.9
W	2	7.3	6.9	15.1	68.7

Table 3.14: Confusion matrix of (Cloud Vision, 2011) system.

Class	D	Ec	S	M	W	A	e-L	T	O	AP
D	72.8	3.1	2.4	3.1	2.9	3.2	2.9	2.4	4.7	2.4
Ec	3.0	73.1	2.4	2.5	2,5	2.7	6.9	2.1	2.1	2.6
S	2.5	2.4	78.6	2.3	2.4	2.2	2.5	2.1	2.9	2.1
M	2.7	2.0	2.7	77.4	2.9	2.6	2.6	2.6	2.1	2.5
W	2.5	2.5	2.2	4.2	72.8	3.6	3.2	2.4	3.8	2.7
A	2.5	2.8	2.1	2.5	3.3	68.7	3.9	3.5	5.7	5.0
e-L	2.5	3.7	2.0	2.5	2.1	2.6	75.6	3.2	3.0	2.7
T	3.1	2.2	2.1	2.9	2.9	4.4	3.4	72.4	4.3	2.3
O	5.5	3.9	4.5	3.7	6.6	5.8	8.2	9.5	48.9	3.4
AP	2.6	2.7	2.5	2.3	2.1	3.2	2.3	2.7	2.8	76.8

Table 3.15: Confusion matrix of the (Bosch, Zisserman, & Muñoz, 2008) classification.

Class	D	Ec	S	M	W	A	e-L	T	O	AP
D	60.4	3.7	1.2	12.0	2.9	6.0	1.2	4.0	1.3	7.3
Ec	0.5	90.5	0.0	0.5	0.3	2.4	5.2	0.6	0.0	0.0
S	0.5	1.0	94.4	2.2	0.9	0.5	0.0	0.2	0.2	0.0
M	1.1	1.8	0.2	90.3	1.7	0.8	0.3	0.4	0.4	3.0
W	0.3	0.6	1.4	5.0	81.8	2.1	1.0	0.3	0.4	7.1
A	5.5	5.1	2.3	8.3	2.8	49.2	3.1	5.4	7.8	10.4
e-L	3.1	13.7	1.7	2.5	0.2	4.0	71.8	1.0	1.0	1.0
T	24.4	10.4	4.6	4.9	3.4	7.4	2.5	32.8	7.2	2.5
O	5.6	1.6	5.6	12.6	5.3	13.9	0.7	13.3	32.8	8.6
AP	18.9	0.6	0.7	1.0	10.6	2.1	0.4	0.3	1.0	64.5

Table 3.16: Confusion matrix of the (Dunlop, 2010) classification.

Class	D	Ec	S	M	W	A	e-L	T	O	AP
D	71.4	5.1	1.0	7.5	4.7	2.1	1.3	1.7	0.8	4.5
Ec	2.1	87.9	0.8	1.2	1.0	2.6	2.0	0.9	0.5	0.9
S	1.0	1.1	90.1	3.1	0.7	0.7	1.2	0.6	1.0	0.6
M	1.4	1.6	1.3	86.3	3.3	1.1	1.4	1.5	0.9	1.0
W	1.2	1.4	1.0	3.9	84.2	0.9	2.6	1.4	0.6	2.9
A	6.5	8.1	7.4	9.1	5.8	46.6	4.1	3.6	3.8	5.1
e-L	3.5	10.5	1.8	1.4	0.5	2.6	77.1	0.9	0.7	0.9
T	15.3	1.7	7.7	5.4	7.1	1.3	4.4	54.0	1.9	1.3
O	9.5	2.2	11.4	14.3	6.3	2.2	1.6	10.1	39.0	3.6
AP	6.8	1.1	2.3	2.8	13.7	0.5	2.9	0.9	0.6	68.6

Table 3.17: Confusion matrix of the (Qin, Shivakumara, Lu, Pal, & Tan, 2016) classification.

Class	D	Ec	S	M	W	A	e-L	T	O	AP
D	53.1	15.3	1.8	7.4	8.8	1.1	2.7	0.5	0.3	9.0
Ec	0.0	97.9	0.2	0.0	0.1	1.3	0.0	0.0	0.0	0.5
S	1.5	0.3	85.6	0.5	4.3	2.3	1.9	2.7	0.4	0.4
M	2.3	1.7	1.3	80.1	3.3	2.7	4.2	0.1	0.1	4.2
W	0.9	0.8	0.8	2.1	93.6	0.0	0.0	0.1	1.1	0.6
A	0.6	13.8	2.8	0.5	2.2	74.7	2.2	0.6	0.5	1.9
e-L	2.8	0.5	3.0	2.9	2.9	1.5	80.5	1.8	0.7	3.4
T	6.8	10.1	13.1	14.9	4.4	2.2	2.8	37.5	2.6	5.6
O	5.6	6.7	13.7	11.8	7.2	14.7	13.7	7.4	7.7	11.6
AP	1.1	1.0	1.7	1.7	1.8	0.7	0.1	0.1	0.0	91.7

Table 3.18 and Table 3.19 report the average R, P, and F-measure of the different text detection approaches on the 5 classes and 10 classes, respectively. For the text detection, the parameters, namely, window size, aspect ratio, window size, aspect ratio for stroke width, the threshold for Bayesian classifier outputs, window size, the threshold for

features vector, aspect ratio for stroke width and the number of sub-blocks are determined, listed in Table 3.18 and Table 3.19.

For Fuzzy-Mass based feature, it is noted from Table 3.18 that the F-measure improves significantly after classification compared to before classification. This validates that the presented approach is effective for enhancing text detection performance.

Table 3.18: Text detection results before and after classification (in %) on data of 5 classes. BC denotes Before Classification, R-Recall, P-Precision, F-Measure and T and w are parameters.

M	(Shivakumara et al., 2012)				(Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010)				(Rong et al., 2014)			
BC	R	P	F	T	R	P	F	T	R	P	F	w
	70.2	52.4	60	0.5	67.2	63.5	65.3	0.06	39.7	27.8	32.7	25
Proposed	69.1	67.2	67.8	0.3	70.9	66.7	68.5	0.2	40.4	38.6	38.7	21
Qin et al., 2016)	61.2	58.7	59.2	0.3	69.1	62.4	65.4	0.2	44.3	36.0	39.5	21
(Bosch et al., 2008)	62.4	61.2	61.6	0.3	60.8	64.1	62.1	0.2	36.4	31.6	33.3	21
Dunlop, 2010	64.4	65.7	64.8	0.3	66.4	62.4	64.2	0.2	33.5	37.8	35.1	21
GOOGLE API)	62.4	63.3	62.4	0.3	62.0	61.9	61.7	0.2	36.1	32.5	34.0	21

For Fuzzy-rough based feature, it is observed from Table 3.19 that the R, P, and F-measure improve significantly compared to prior to classification for the text detection methods. This shows that classification is effective especially when frames have large variations in background and foreground complexities for increasing the performances of text detection techniques. At the same time, when text detection performance is compared for the proposed classification with the existing classification methods, most of the text detection methods score highest F-measure for the proposed classification method than the existing methods. The assumption is that text detection methods might score good results if the classification methods classify frames correctly without many misclassification errors. Therefore, since the proposed classification obtains the best classification rate, most of the text detection methods perform better for the proposed classification. However, Li et al.'s (Huiping Li et al., 2000) technique accomplishes the best F-measure for Bosch et al.'s (Bosch et al., 2008) classification, Rong et al.'s (Rong et al., 2014) algorithm achieves the best F-measure for Dunlop's (Dunlop, 2010) classification, Shivakumara et al. (Shivakumara et al., 2012) obtains the best F-measure for Qin et al.'s (Qin et al., 2016) classification, and Zhao et al. (X. Zhao et al., 2011) attains the best F-measure for GOOGLE API classification (Cloud Vision, 2011).

Table 3.19: Text detection performance of the different existing methods prior to classification and after classification for proposed and existing classification methods on data of 10 classes. PC denotes "Prior to classification" and AC denotes "After classification".

Text Detection Methods	Evaluation Metric	PC	AC				
			Proposed	(Bosch et al., 2008)	(Dunlop, 2010)	(Qin et al., 2016)	GOOGLE API)
(Khare, Shivakumara, & Raveendran, 2015)	R	32.2	50.3	43.3	50.1	53.3	40.3
	P	40.5	55.7	53.2	53.2	47.6	59.7
	F	35.8	52.6	47.7	51.6	50.2	48.1
(Yin et al., 2014)	R	42.3	55	52.6	45.1	47.4	50.4
	P	48.4	58.6	46.8	47.7	42.4	53.5
	F	45.1	56.5	49.5	46.3	44.7	51.9
(Rong et al., 2014)	R	33.2	45.5	51.3	55.1	44.1	42.5
	P	37.2	55.6	50.4	52.9	50.4	57.2
	F	35	50	50.8	52.9	47	48.7
(Mosleh et al., 2013)	R	35.4	52	47.2	56.1	52.8	54.2
	P	44.1	56.8	56.2	46.1	47.2	50.2
	F	39.2	53.7	51.3	50.6	49.8	52.1
(Shivakumara et al., 2012)	R	33.6	51.3	55.1	55.8	52.5	53.2
	P	42.7	51.4	49.9	52.8	58.4	45.4
	F	37.6	50.4	52.3	54.2	55.2	48.9
(X. Zhao et al., 2011)	R	32.6	45.5	40.2	42.9	47.1	43.4
	P	39.6	52.6	47.3	48.8	52.3	53.5
	F	35.7	47.8	43.4	45.6	59.5	53.5
(Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010)	R	38.4	50.9	51.8	46.9	43.3	53.2
	P	27.2	53.1	38.9	55.2	50.1	40.3
	F	31.8	51.6	44.4	50.7	46.5	45.8

Table 3.19: Continued

Text Detection Methods	Evaluation Metric	PC	AC				
			Proposed	(Bosch et al., 2008)	(Dunlop, 2010)	(Qin et al., 2016)	GOOGLE API)
(Epshtein et al., 2010)	R	31.6	53.1	54.7	50.2	46.2	50.2
	P	35.8	59	53.3	52.8	45.8	57
	F	33.8	55.7	53.9	51.4	45.9	53.3
(Huiping Li et al., 2000)	R	31.3	41.9	46.2	40.1	45.7	42.8
	P	37.4	53	54.9	56.3	48.2	51.2
	F	34	46.4	50.1	46.8	46.9	46.6

Table 3.20 and Table 3.21 shows that the RR after classification improve greatly for all the binarization methods of all the classification methods compared to the RR of prior to classification. This is possible because of tuning the parameters based on samples of each class by considering advantage of classification step. When the recognition accuracy of the presented classification is compared with the existing classifications, most binarization algorithms accomplish the best RR for the proposed classification compared to the existing classification. However, Milyaev et al.'s method (Milyaev et al., 2013) achieves the best recognition rate for GOOGLE API classification. The reason for the poor recognition performances by different binarization methods for the existing classification is the same as the reason discussed in the previous section.

Table 3.20: Average recognition rate (%) of the different binarization methods for the proposed and existing classification methods on data of 5 classes.

Classification Methods	(S. Roy et al., 2015)		(Howe, 2013)		(Milyaev et al., 2013)	
Before classification						
Proposed	17.9	0.05	18.4	10	16.6	0.4
(Bosch et al., 2008)	17.9	0.05	16.6	0.4	18.4	10
(Dunlop, 2010)	17.9	0.05	16.6	0.4	18.4	10
(Qin et al., 2016)	17.9	0.05	16.6	0.4	18.4	10
(Engine, 2008)	17.9	0.05	18.4	10	16.6	0.4
After classification						
Proposed	25.7	0.22	35.5	0.27	18.4	10
(Bosch et al., 2008)	21.6	0.2	30.8	0.2	29.0	9.4
(Dunlop, 2010)	23.0	0.2	32.3	0.2	32.8	9.4
(Qin et al., 2016)	23.2	0.2	33.2	0.2	30.7	9
(Engine, 2008)	21.8	0.2	29.8	0.2	23.4	9.4

Table 3.21: Average recognition rate of the different binarization methods for the proposed and existing classification methods on data of 10 classes.

Methods	(S. Roy et al., 2015)	(Su & Lu, 2014)	(Howe, 2013)	(Milyaev et al., 2013)
Prior to classification	15.78	12.21	13.14	10.31
Classification methods	Before and after classification			
Proposed	32.3	40.1	37.1	37.2
(Bosch et al., 2008)	23.2	28.2	24.7	35.8
(Dunlop, 2010)	28.9	23	22.2	30.4
(Qin et al., 2016)	30.9	27	20.2	19.4
(Engine, 2008)	31.8	36.7	35.2	39.3

In order to test the generic nature and robustness of the presented approach by text detection and recognition, 5 new classes are chosen, namely, Recipes of Cooking (RC) which contains text as Animal Planet, Craft Making (CM) which contains caption text as Animal Planet, Indian Classical Musical Concert (ICMC) which contains texts as in Sports, Outlet, Defense, Teleshopping (TS) which contains caption texts as in Sports, Animal Planet, and Yoga (Y) which contain caption texts as in Animal Planet. The sample images of new classes with texts are shown in Figure 3.20.

For experimentation of these new classes, the same setup which has been used for the 10 classes, is used. Table 3.22 and Table 3.23 report the quantitative results of text detection and binarization methods on the data of 5 new classes. With the same parameter setup, the results reported in Table 3.22 and Table 3.23 show that the text detection and recognition performance improves significantly after classification than those of prior to classification with the similar conclusion derived for the data of 10 classes. Furthermore, it is noted from the results of 10 classes and 5 new classes, the detection and recognition rates report almost similar patterns after classification.

Table 3.22: Text detection performance of the different existing methods prior to classification and after classification for proposed and existing classification methods on new 5 classes. PC denotes "Prior to classification" and AC denotes "After classification".

Text Detection Method	Evaluation Metric	PC	AC				
			Proposed	(Bosch et al., 2008)	Dunlop, 2010)	(Qin et al., 2016)	GOOGLE API)
(Khare, Shivakumara, & Raveendran, 2015)	R	34.7	48.7	45.2	54.1	51.6	42.1
	P	45.4	54	52.5	50.9	46.2	45.4
	F	40	51.3	48.5	52.4	48.7	43.6
(Yin et al., 2014)	R	46.1	54	50.8	52.6	35.4	51.2
	P	50.2	56.5	42.5	47.4	39.8	49.1
	F	48.1	55.2	46.2	47	38.6	50.1
(Rong et al., 2014)	R	35.8	45.2	43.5	44.1	47.3	47.2
	P	38.9	50.9	51.8	41.5	48.5	54.1
	F	37.3	48.04	42.9	43.2	44.7	50.4
(Mosleh et al., 2013)	R	37.9	54.9	47.3	50.3	53.1	52.1
	P	48.2	58.8	48.5	53.1	45.4	57.5
	F	43	56.9	44.7	46	47.2	54.6

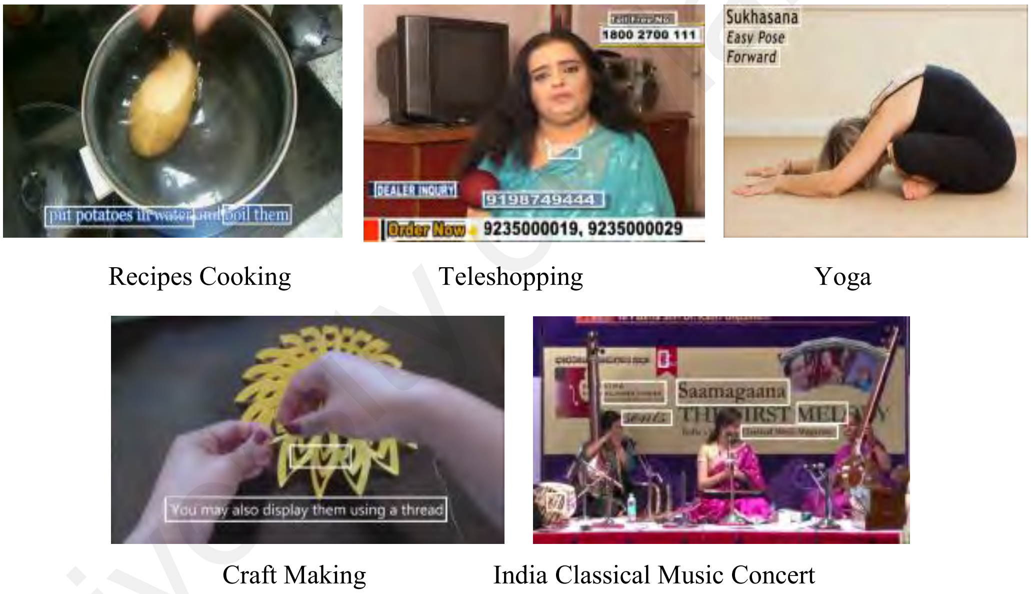
Table 3.22: Continued

Text Detection Method	Evaluation Metric	PC	AC				
			Proposed	(Bosch et al., 2008)	Dunlop, 2010)	(Qin et al., 2016)	GOOGLE API)
(Shivakumara et al., 2012)	R	38.7	55.9	50.4	54.3	57.2	47.6
	P	42.3	59.5	57.1	57.1	44.5	54.2
	F	40.2	57.7	46.1	47.6	48.7	50.6
(X. Zhao et al., 2011)	R	39.7	59.9	41.5	44.1	48.6	47.2
	P	42.8	55	43.8	50.2	44.1	57.8
	F	41.2	52.4	41.9	43.2	45.3	51.9
(Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010)	R	42.7	53.2	52.5	50.1	52.5	56.1
	P	35.3	56.4	48.4	44.5	48.1	52.2
	F	39	54.8	46.9	45.9	46.9	54.1
(Epshtein et al., 2010)	R	37.4	49.1	50.2	48.5	51.5	52.6
	P	39.9	54.4	43.4	47.5	38.2	54.2
	F	38.6	51.7	46	45.3	46.5	53.3
(Huiping Li et al., 2000)	R	36.1	49.2	48.3	41.7	48.7	42.4
	P	39.4	51.4	41.1	42.8	44.5	50.5
	F	37.7	49.9	45.2	42	45.3	46

Table 3.23: Average recognition rate (%) of the different binarization methods for the proposed and existing classification methods on data of 5 new classes.

Methods	(S. Roy et al., 2015)	(Su & Lu, 2014)	(Howe, 2013)	(Milyaev et al., 2013)
Prior to classification	19.7	16.7	18.3	17.9
Classification methods	Before and after classification			
Proposed	35.1	33.1	27.8	36.6
(Bosch et al., 2008)	23.1	34.7	21.4	32.9
(Dunlop, 2010)	24.4	22.1	24.5	23.1
(Qin et al., 2016)	20.8	23.4	21.2	30.4
(Engine, 2008)	30.5	33.1	27.8	32.9

In summary, the presented approach has the ability to extend to a number of new classes, and the performance of classification is independent of the content of frames. Hence, the proposed method is generic and consistent to different classes or different contents of frames. This is because of the use of the flexible rough-fuzzy combination, covariance-correlation for intra, inter planes and temporal information. Since the aim is to validate the efficacy of the proposed categorization method, parameters of the text detection and binarization methods are tuned. The improved results after classification are not high as plain document analysis accuracy which usually has more than 90% accuracy. However, this work shows the direction that by considering the advantage of classification, one can modify the existing methods or develop new methods of text detection and recognition according to the complexity of individual classes for achieving better accuracy.


put potatoes in water and boil them
Toll Free No.
1800 2700 111
DEALER INQUIRY
9198749444
Order Now 9235000019, 9235000029
Sukhasana
Easy Pose
Forward
You may also display them using a thread
Saamagaana
THE FIRST MELODY

Recipes Cooking Teleshopping Yoga

Craft Making India Classical Music Concert

Figure 3.20: Samples of video frames of new 5 classes with text detection by (Yin et al., 2014).

3.5 Summary

Overall, in this chapter, two approaches are presented based on Fuzzy for video scene classification. The first approach explores a combination of Fuzzy-mass based feature based on spatial proximity to classify very primitive types of edge, whereas the second approach explores rough set and fuzzy logic combination to classify complex edges. Both the approaches extract local/intra and global/inter, which helps to encode unique relationship for each video class type. Temporal information is used to increase the discriminative power of feature extraction. Experimental results on classification validate that the discussed approach is superior for different scene type videos compared to the existing techniques. In addition, the usefulness and efficacy of the proposed classification are validated by text detection and recognition experiments.

CHAPTER 4: FRACTIONAL POISSON ENHANCEMENT MODEL FOR TEXT DETECTION AND RECOGNITION IN VIDEO FRAMES

4.1 Background

In the previous chapter, Fuzzy based video scene classification methods have been discussed. Generally video has low contrast and low resolution, therefore, it is difficult to achieve better recognition rate. One common enhancement technique is Laplacian-based methods. However, the Laplacian operation introduces noise during enhancement. Therefore, this chapter focuses on removing noise created by Laplacian operation and enhancing low contrast text in the video as well.

4.2 Overview of Fractional Poisson Model

Fractional calculus and its applications are widely used in engineering and sciences. Moreover, fractional differentiation is considered as an outstanding mathematical tool to describe the dynamic behavior of various materials and systems (Jalab & Ibrahim, 2015; C. Zhou, Yan, Tao, & Lui, 2012). For the last 50 years, numerous operators of fractional calculus have been developed, such as Erdélyi-Kober, Grünwald-Letnikov, Caputo, Riemann-Liouville, and Weyl-Riesz operators (Jalab & Ibrahim, 2015; C. Zhou et al., 2012). Fractional calculus has received significant attention in image processing domain, for example, image denoising, and image texture improvement. Fractional calculus operators generate high levels of stable invulnerability against various noises (Jalab & Ibrahim, 2015; C. Zhou et al., 2012).

The logic behind the image enhancement base on fractional Poisson is that the images obtained by Laplacian operation will introduce an image with insignificant changes in gray level. However, fractional Poisson maintains high-frequency details while enhancing low-frequency details due to its non-linearity nature. This observation motivates to utilize the fractional calculus to increase the quality of images obtained by Laplacian operation.

Generally, image enhancement approaches are employed on the original image so that the enhanced image shows better result compared to the original image. A generalized model is proposed using fractional Poisson to enhance the quality of the image, obtained by Laplacian operation. In our work, image improvement or enhancement indicates the changes in original pixels values to achieve better contrasts between the target and their surrounding area.

4.3 A Model for Video Image Enhancement

Our investigation is based on the Riemann–Liouville fractional differential operator of the order 0 < α <1 (Podlubny, 1998)

$$D^{\alpha} f(t) = \frac{d}{dt}\int_{a}^{t}\frac{(t-\tau)^{-\alpha}}{\Gamma(1-\alpha)} f(\tau)d\tau, \qquad (4.1)$$

The following equation corresponds to the fractional integral operator for a continuous function *f*(t) of the order α > 0:

$$I_{a}^{\alpha} f(t) = \int_{a}^{t}\frac{(t-\tau)^{\alpha-1}}{\Gamma(\alpha)} f(\tau)d\tau, \qquad (4.2)$$

A fractional non-Markov Poisson stochastic process is implemented using a Riemann–Liouville fractional differential operator, and Kolmogorov–Feller equation (Laskin, 2003). For estimating a probability distribution function *P*(x,s), the fractional Kolmogorov–Feller equation is defined by

$$\frac{\partial P(x,s)}{\partial s} = \int_{-\infty}^{\infty} dy\, \omega(y)[P(x-y,s) - P(x,s)],$$

$$P(x,0) = \delta(x),$$

where ω corresponds to the probability density of length *y* and *s* refers the time steps of order α. The randomness of step length is distributed according to ω. Therefore, fractional Poissonian distribution is defined as follows:

$$\Psi(s) = \frac{\sin \pi\alpha}{\pi} \int_0^{\infty} \frac{e^{-\rho s} d\rho}{2\cos(\pi\alpha)+\rho^{\alpha}+\rho^{-\alpha}}, \quad 0 < \alpha \leq 1,$$

Let *P*(*n*,*r*) be the probability of *n* items in position *r*. The probability *P* satisfies the normalizing condition when $\sum_{n=0}^{\infty} P(n,r) = 1$.

In general, a specific type of the fractional Kolmogorov–Feller equation (Laskin, 2003) defines the probability $P_{\alpha}(x,s)$:

$$P_{\alpha}(n,r) = \frac{(r^{\alpha}\,\bar{n})^n}{n!} \sum_{k=0}^{\infty} \frac{(k+n)!}{k!} \frac{(-r^{\alpha}\bar{n})^k}{\Gamma(\alpha(k+n)+1)}, 0 < \alpha \leq 1, \quad (4.3)$$

Therefore, the mean $\overline{\mathrm{n}_{\alpha}}$ of the fractional Poisson process has been estimated straightforward as follows:

$$\overline{n_{\alpha}} \;\; = \sum_{n=0}^{\infty} n\, P_{\alpha}(n,r) = \frac{\bar{n}\, r^{\alpha}}{\Gamma(\alpha+1)}, \quad (4.4)$$

where $\bar{n}$ is the mean value of the image, which represents the average of all pixels of an image, α is the fractional power, while *r* is a tuning parameter used for enhancing image contrast to a further level. The mean $\overline{n_{\alpha}}$ is modified in equation 4.4 to be the fractional mean $\overline{m_{\alpha}}$ of Laplacian image. The mean has been modified to use it as intensity transform for increasing the dynamic range of the gray level, which is as given by:

$$\overline{m_{\alpha}} \;\; = \sum_{n=0}^{\infty} n\, P_{\alpha}(\overline{n_{\alpha}}, r) \approx \frac{(\overline{n_{\alpha}}\; r)^{\alpha}}{\Gamma(\alpha+1)}, \quad (4.5)$$

In the context of image processing, an enhanced image is obtained as per the following formula:

$$Ie = \frac{(\overline{m_\alpha}\ \ r)^\alpha}{\Gamma(\alpha+1)} * I \ \ , \tag{4.6}$$

where Ie is the generated enhanced image, * is denoted by the convolution product, and I is the Laplacian image. Here, the contrast enhancement depends on the fractional mean $\overline{m_\alpha}$, obtained from the Laplacian image. In this, $\overline{m_\alpha}$ is made adaptive, i.e., $\overline{m_\alpha}$ is of a lower value for dark pixels and at the same time $\overline{m_\alpha}$ is of a higher value for bright pixels. The fractional mean depends on image content as well as the value of α. Here, the fractional mean is introduced to enhance the dynamic range of the gray level adaptively.

In general, image contrast is defined by the discrepancy between the visual properties of target objects and their surrounding in the image.

For the proposed fractional image enhancement algorithm, the steps are as follows:

i. Define the values of the fractional parameter α with the range of $0< \alpha \leq 1$ and r with the range of $0< r \leq 1$.
ii. Estimate the mean value of the Laplacian image.
iii. Calculate the fractional mean using equation (4.5).
iv. Enhance the Laplacian image using equation (4.6).

These two parameters α, and r contribute significantly to increase the image contrast. From equation (4.5), it can be seen that the fractional mean $\overline{m_\alpha}$ and the tuning parameter r, have been raised by the power of α. Therefore, Laplacian image is affected by the distinct values of α and r. Figure 4.1 illustrates the PSNR behavior of the generated enhanced image for multiple values of α and r. The dynamic range of dark pixels is extended by lower values of α, which corresponds to a small PSNR and the dynamic range of bright pixels is extended by larger values of α, which corresponds to a dramatic decrease in PSNR. In its simplest form, the optimal values of α and r are chosen manually.

From Figure 4.1, $\alpha = 0.6$ is chosen with the tune parameter r=0.04 (Dash-dot line blue color).

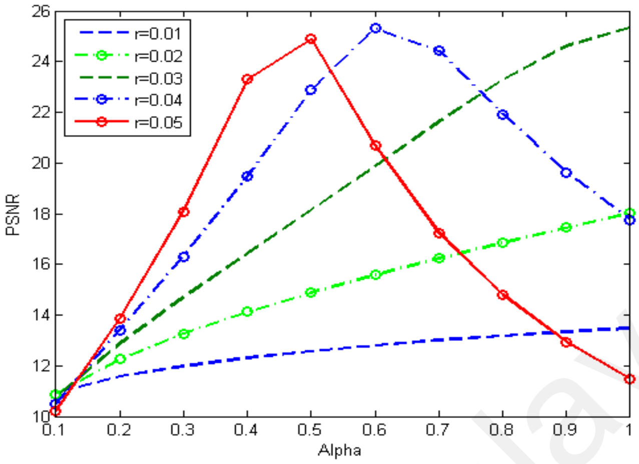


Figure 4.1: The behavior of PSNR for the values of α, and r.

4.4 Experimental Results and Comparative Study

Section 4.4 is organized as follows. Section 4.4.1 gives details description the datasets and evaluation metrics used for proposed approach. The evaluation metrics have been presented in terms of quality, text detection experiments, and recognition. Section 4.4.2 presents the experiments of measuring the quality of enhancement method and Section 4.4.3 discusses the experiments for validating the effectiveness of proposed enhanced result. A comparative study is performed on recent text detection and recognition methods to validate the efficiency of the presented approach in section 4.4.4.

4.4.1 Datasets and Evaluation

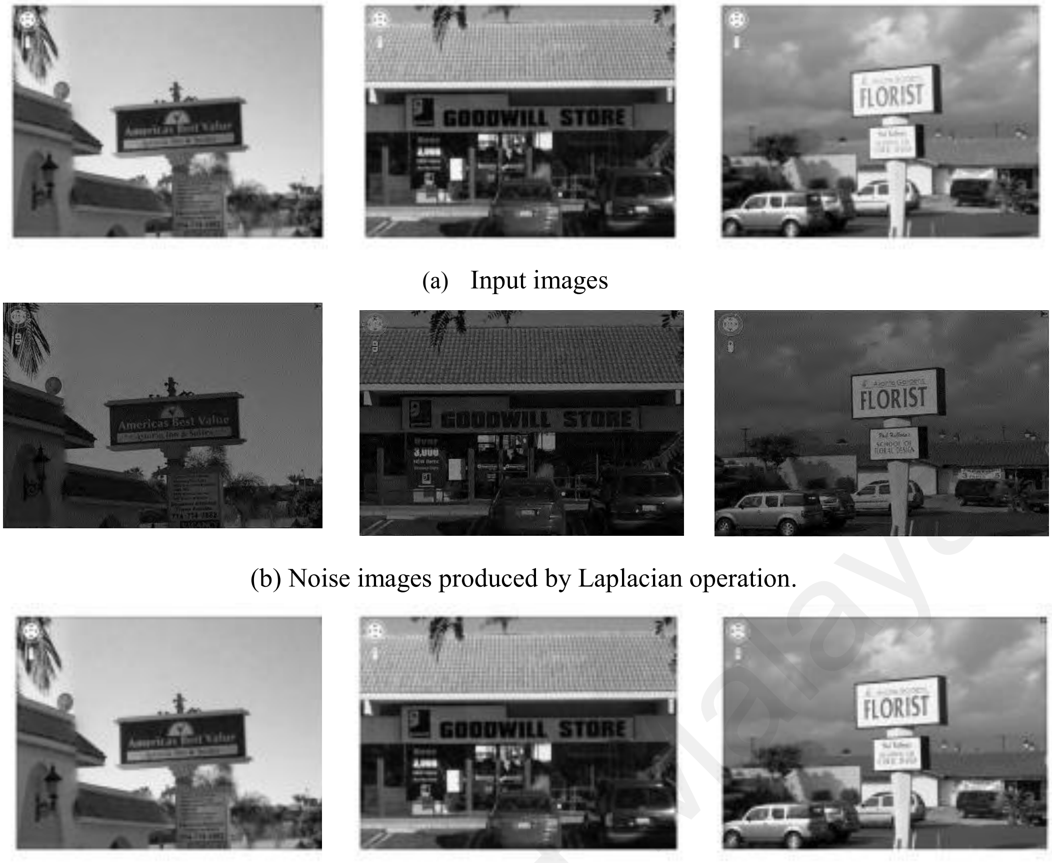
Standard datasets, namely: (i) the ICDAR 2013 video dataset (Karatzas et al., 2013) which includes texts with low resolution, complex background, different fonts or font sizes and different orientations, (ii) the ICDAR 2013 scene dataset (Karatzas et al., 2013) which includes mainly horizontal texts with high contrast in complex background, (iii) the Street View Data (SVT) (K. Wang & Belongie, 2010) which includes high resolution, complex background, small fonts, and distorted texts and (iv) the MSRA data (Yao et al.,

2012) which includes high resolution, complex background and arbitrarily oriented text lines have been used. In total, 6366 text images are used for experimentation in this work. The main advantage of these datasets is that all of them provide ground truth for calculating measures without involving human intervention. The quality of the enhanced images is measured by two standard measures, namely, Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) (Jalab & Ibrahim, 2015; C. Zhou et al., 2012). The performance of the introduced model is compared with the existing enhancement works, such as, such as Histogram Equalization (HE), Contrast-Limited Adaptive Histogram Equalization (CLAHE), which are employed to enhance the contrast of image by altering the image intensities of grayscale images, and Adjust Intensity Values to Specified Range (AIV) which is operated to increase the image contrast by mapping the intensity values of the input grayscale image to new values. The reason to choose these three techniques is that majority of the enhancement approaches proposed in literature directly or indirectly used these techniques as a basis to develop methods (Jalab & Ibrahim, 2015; C. Zhou et al., 2012). Therefore, the proposed model is compared with these bases to show that the proposed model is effective and accurate in this work.

Since datasets do not have ground truth for noise images (Laplacian images), experiments are conducted on the standard datasets, explained above. It is expected that the performance of text identification and recognition results should be better in case of enhanced images compared to the input and Laplacian images. Therefore, the existing text extraction methods, namely, Epshtein et al. method (Epshtein et al., 2010) which uses SWT for scene images, Shivakumara et al. approach (Palaiahnaktoe Shivakumara et al., 2010) which integrates wavelet and color features in video, Shivakumara et al. method (Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010) which uses Fourier transform and RGB color in video, Shivakumara et al. method (Shivakumara et al., 2012) which proposes Bayesian theory in video images, and Rong et al. method (Rong et al., 2014) which employs CC analysis and a classifier in scene images, have been implemented. Similarly, for recognition experiments, binarization methods, namely, Wolf et al. method (Wolf et al., 2002) which uses edge and gradient information for binarization, Roy et al. method (S. Roy et al., 2012) which uses gradient and wavelet fusion for video text lines, Chattopadhyay et al. method (Chattopadhyay et al., 2013b) which selects automatically binarization methods for different portions of a document, Moghaddm and Cheriet's method (Moghaddam & Cheriet, 2010) which proposes adaptive binarization for document images, and Howe's method (Howe, 2013) which proposes to binarize images by tuning parameters automatically, have been implemented. Recall, precision and f-measure are used for evaluating the performances of the text detection methods as these are the standard measures for text detection. The recognition rate at character level given by the OCR engine is used for evaluating the binarization methods.

4.4.2 Experiments for Measuring the Quality of the Enhanced Image

The result of the presented model can be visualized in Figure 4.2, where (a) shows different input images, (b) denotes the noisy/corrupted images after applying Laplacian operation, and (c) presents enhanced images obtained by the proposed model. Figure 4.2 depicts that the enhanced images become brighter compared to the (a) and (b). Moreover, there is no noise in the enhanced image as in the Laplacian image. This helps the model to accomplish better result in text extraction and recognition approach.



(a) Input images

(b) Noise images produced by Laplacian operation.

(c) Enhanced images by the proposed model.

Figure 4.2: Sample qualitative results of the proposed model.

The quality measures can be defined as in equation (4.7) and equation (4.8). PSNR is estimated by calculating the MSE between Laplacian image (*C*) and the original image (*O*) (Jalab & Ibrahim, 2015; C. Zhou et al., 2012) for each pixel. Therefore PSNR is defined as follows:

$$\mathrm{PSNR} = 10\log\frac{\max(C,O)^2}{\mathrm{MSE}} \tag{4.7}$$

$$\mathrm{MSE} = \frac{1}{\mathrm{MN}}\sum_{\mathrm{i=1}}^{\mathrm{M}}\sum_{\mathrm{j=1}}^{\mathrm{N}}(C(i,j) - O(i,j))^2 \tag{4.8}$$

where *max* is the maximum possible pixel value of the image. For example, the *max* is 255 for a grayscale image. PSNR is measured in decibels and it is inversely proportional to MSE. A larger value of PSNR indicates more similarity between the images.

Similarly, SSIM can be defined as in equation (4.9). The SSIM is introduced in order to find about all the ways to compare the structures of the original and the Laplacian images. This metric is defined as follows:

$$\mathrm{SSIM}(x,y)=[\,l(x,y)]^{\alpha}.\,[c(x,y)]^{\beta}.\,[s(x,y)]^{\gamma} \qquad (9)$$

Where l is the luminance comparison function, c is the contrast comparison function, and s is the structure comparison function. The parameters α, β, and γ are used to adjust the relative importance of the three components (Z. Wang, Bovik, Sheikh, & Simoncelli, 2004).

The two quality measures are calculated for different combinations, such as an enhanced image with a Laplacian image, an original image with a Laplacian image, and an enhanced image with an original image because the ground truth is not available for the input Laplacian images. Ideally, PSNR and SSIM should give high scores for an enhanced image with its original image, while for an enhanced image with its Laplacian image, the measures should give lower scores compared to those of the original image with the Laplacian image. Table 4.1 reports the quantitative accuracies of the presented approach. It is noticed in Table 4.1 that both the measures score higher values for the enhanced image with the original image (input image), and lower score for the enhanced image with the Laplacian image. This conclusion can be verified with the scores of the original image (input image) with the Laplacian image because for this combination, the scores are neither higher than those of Enhanced-Original (input), nor lower than those of Enhanced- Laplacian. Therefore, presented enhancement model is promising.

Table 4.1: Average PSNR and SSIM for enhanced images with noisy and input images

Dataset	Number of images	PSNR			SSIM		
		Enhanced - Laplacian	Original-Laplacian	Enhanced – Original (proposed)	Enhanced –Laplacian	Original-Laplacian	Enhanced–Original (proposed)
ICDAR 2013 Video	5487	11.95	13.15	24.26	0.74	0.76	0.95
ICDAR 2013 Scene	229	12.10	13.39	22.82	0.74	0.76	0.91
SVT	350	12.9	15.15	24.56	0.74	0.79	0.96
MSRA	300	11.7	12.7	26.99	0.74	0.76	0.98
Phos image set	100	8.39	8.84	18.26	0.70	0.65	0.89
Mean	Total = 6366	11.41	12.64	23.37	0.73	0.74	0.93

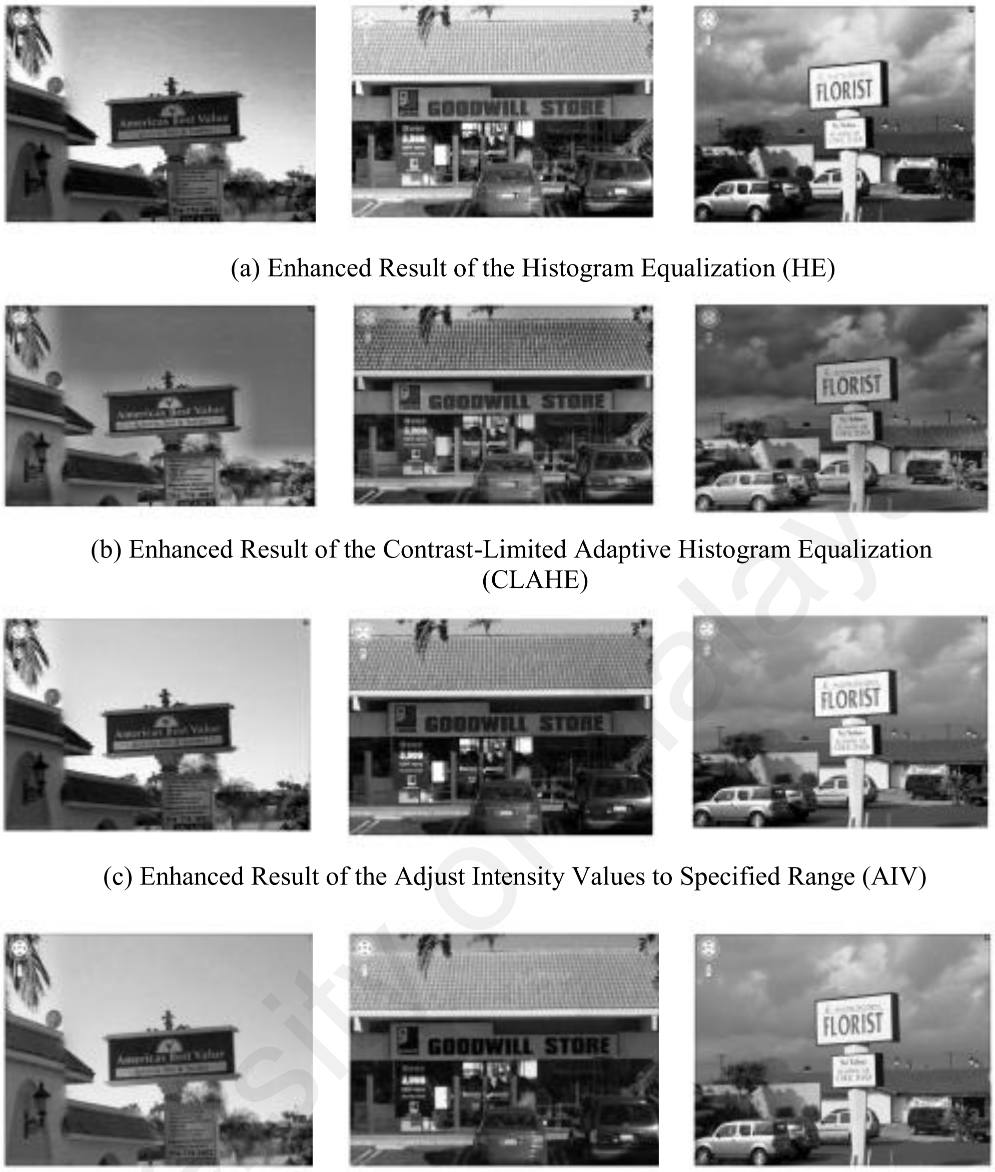
4.4.3 Experiments for validating the effectiveness of the Enhanced Results

In this work, the noises generated by Laplacian operation is considered as a case study for building the enhancement model rather than adding noises to the input image manually as denoising methods work in the literature. As a result, corrupted data do not have ground truth for calculating quality measures. An input image is used as the original image for quality measure calculation. Therefore, to prove the efficiency of the introduced model, experiments are done on extraction and recognition of the texts in the input image, the Laplacian image, and the enhanced image. Moreover, to show that text identification and recognition methods give better accuracies for enhanced images compared to input and Laplacian images.

4.4.4 Comparative Study and Discussion

Figure 4.3 depicts some sample qualitative results of the proposed and existing methods. In Figure 4.3, (a) gives the results of HE, (b) shows the results of the CLAHE,

(c) illustrates the results of AIV, and (d) shows the results of the proposed model. From the results in Figure 4.3, visually it is hard to notice any difference. However, the quantitative differences between the proposed and existing works can be noticed in Table 4.2. It can be seen from Table 4.2 that the PSNR results of the proposed algorithm outperform those of the existing methods, namely, HE, CLAHE, and AIV, for all the datasets considered for experimentation. This is due to the capability of the proposed algorithm to adopt suitable image enhancement parameters automatically by applying the fractional mean $\overline{m_\alpha}$ of the input Laplacian image as an intensity transform to increase the dynamic range of the gray levels of the Laplacian image. The comparison verifies that the proposed model can be effectively applied to enhance contrast, and thus achieves a better image enhancement performance than other methods.



(a) Enhanced Result of the Histogram Equalization (HE)

(b) Enhanced Result of the Contrast-Limited Adaptive Histogram Equalization (CLAHE)

(c) Enhanced Result of the Adjust Intensity Values to Specified Range (AIV)

(d) Enhanced Results of the Proposed Model

Figure 4.3: Sample qualitative results of the proposed model and existing techniques.

Table 4.2: Quality measures of the proposed and existing models (PSNR and SSIM are calculated for Enhanced images with Original (input) images).

Dataset	HE		CLAHE		AIV		Proposed Model	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ICDAR 2013 Video	18.1	0.73	18.1	0.78	17.7	0.78	24.3	0.95
ICDAR 2013 Scene	14.8	0.63	14.8	0.63	14.8	0.63	22.8	0.91
SVT	15.8	0.72	15.8	0.72	15.8	0.72	24.5	0.96

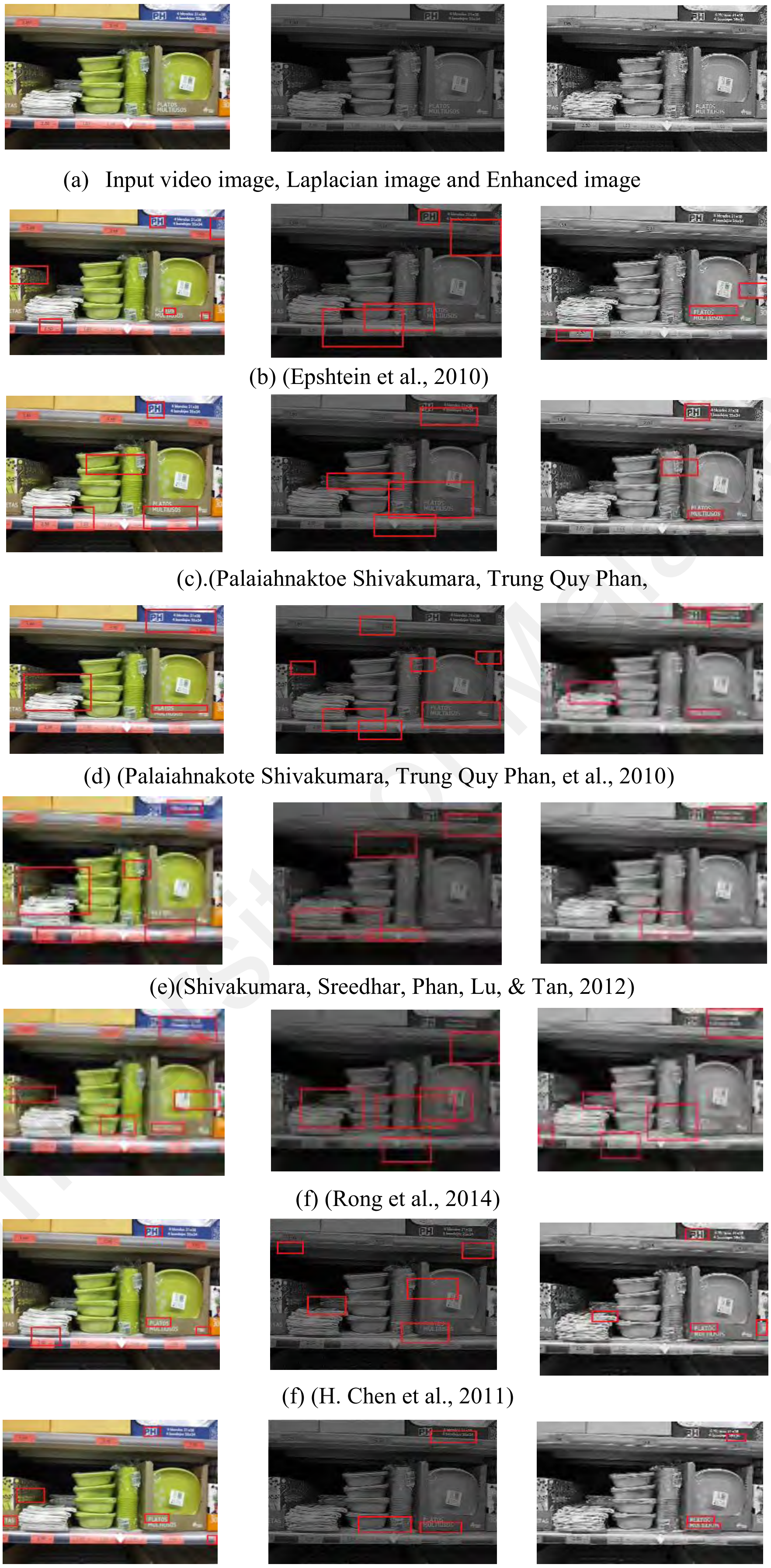
Table 4.2: Continued

Dataset	HE		CLAHE		AIV		Proposed Model	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MSRA	16.6	0.75	16.6	0.75	16.6	0.75	26.9	0.98
Phos image database	12.6	0.64	11.1	0.80	11.1	0.80	18.2	0.98
Average	15.58	0.694	15.28	0.736	15.2	0.736	23.34	0.956

The text detection results for the datasets, namely, ICDAR 2013 video, ICDAR 2013 scene, Street View Data and MSRA data, are respectively shown in Table 4.3-Table 4.6. Different text detection methods have been used for validating the results of the proposed enhancement model. It is seen from Table 4.3-Table 4.6 that the text detection results give more false positive for the input images, worst results for the Laplacian images due to noise introduction but better results for the enhanced images. The same can be observed from Figure 4.3-Figure 4.7. This infers that the noises created by Laplacian operation affect the text detection performance significantly. Therefore, it can be concluded that the detection results improve significantly for the enhanced images.

Table 4.3: Text detection before and after enhancement on ICDAR 2013 Video data.

Methods	Before Enhancement						After Enhancement		
	Original			Laplacian			Enhanced		
	R	P	F	R	P	F	R	P	F
(Epshtein et al., 2010)	57.23	50.17	53.46	38.15	25.2	30.35	53.25	63.1	57.75
(Palaiahnaktoe Shivakumara et al., 2010)]	38.12	62.32	47.30	39.46	28.12	32.83	46.23	61.13	52.64
(Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010)	65.1	42.62	51.51	73.4	26.21	38.62	65.12	52.12	57.89
Shivakumara, Sreedhar, Phan, Lu, & Tan, 2012)	55.21	29.19	38.18	45.11	21.87	29.45	50.21	34.22	40.70
(Rong et al., 2014)	36.0	25.12	29.59	30.12	13.10	18.25	45.0	22.5	30.0
(H. Chen et al., 2011)	51.16	57.48	54.13	31.41	35.12	33.16	55.83	59.24	57.48
(Yin et al., 2014)	52.73	64.31	57.94	48.21	52.6	50.30	57.27	64.3	60.58



(a) Input video image, Laplacian image and Enhanced image

(b) (Epshtein et al., 2010)

(c).(Palaiahnaktoe Shivakumara, Trung Quy Phan,

(d) (Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010)

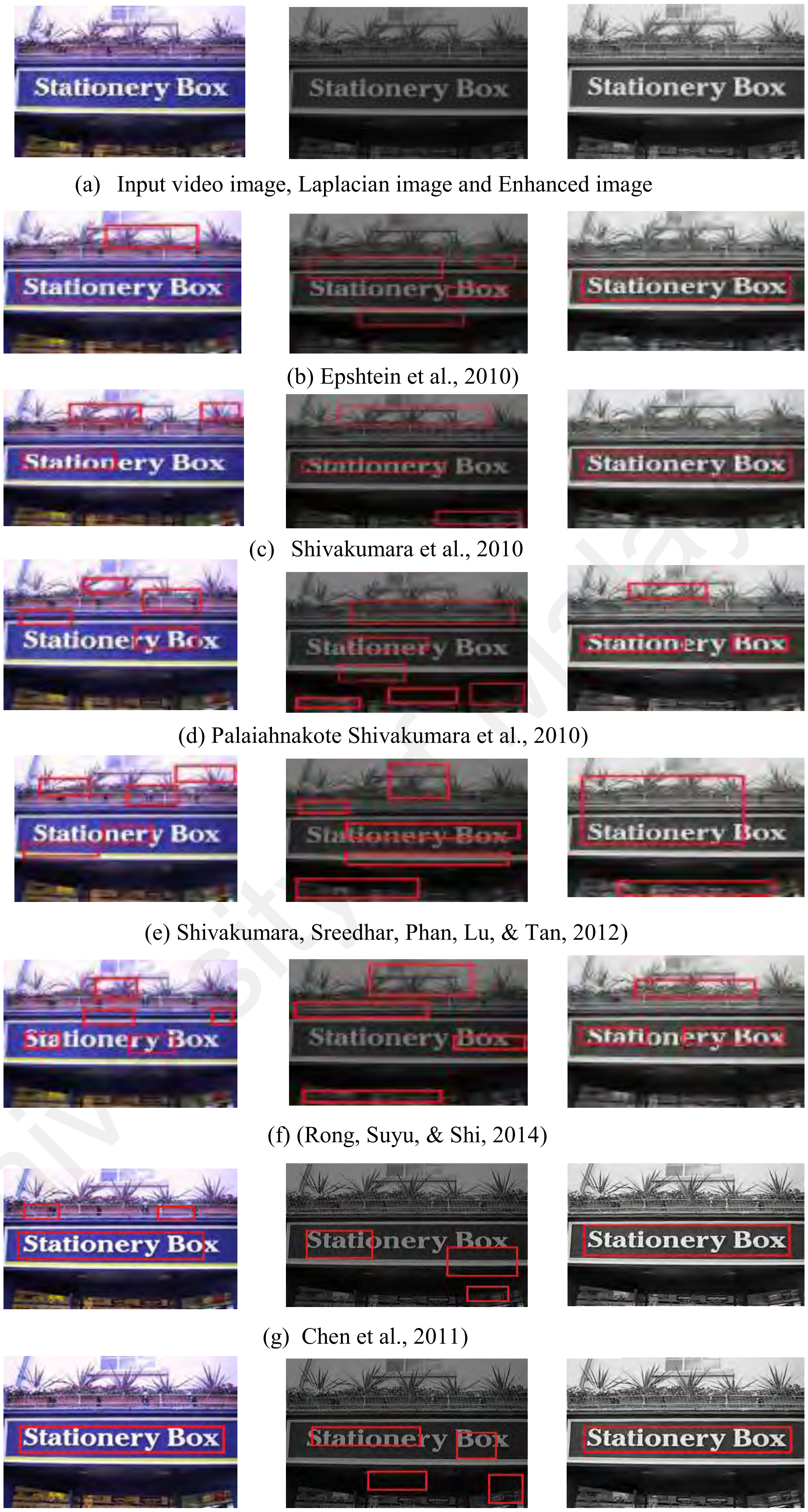
(e)(Shivakumara, Sreedhar, Phan, Lu, & Tan, 2012)

(f) (Rong et al., 2014)

(f) (H. Chen et al., 2011)

(h) (Yin, Yin, Huang, & Hao, 2014)

Figure 4.4: Sample qualitative results of the different text detection methods on input, Laplacian and enhanced images for ICDAR 2013 video frames.



(a) Input video image, Laplacian image and Enhanced image

(b) Epshtein et al., 2010)

(c) Shivakumara et al., 2010

(d) Palaiahnakote Shivakumara et al., 2010)

(e) Shivakumara, Sreedhar, Phan, Lu, & Tan, 2012)

(f) (Rong, Suyu, & Shi, 2014)

(g) Chen et al., 2011)

(h) (Yin et al., 2014)

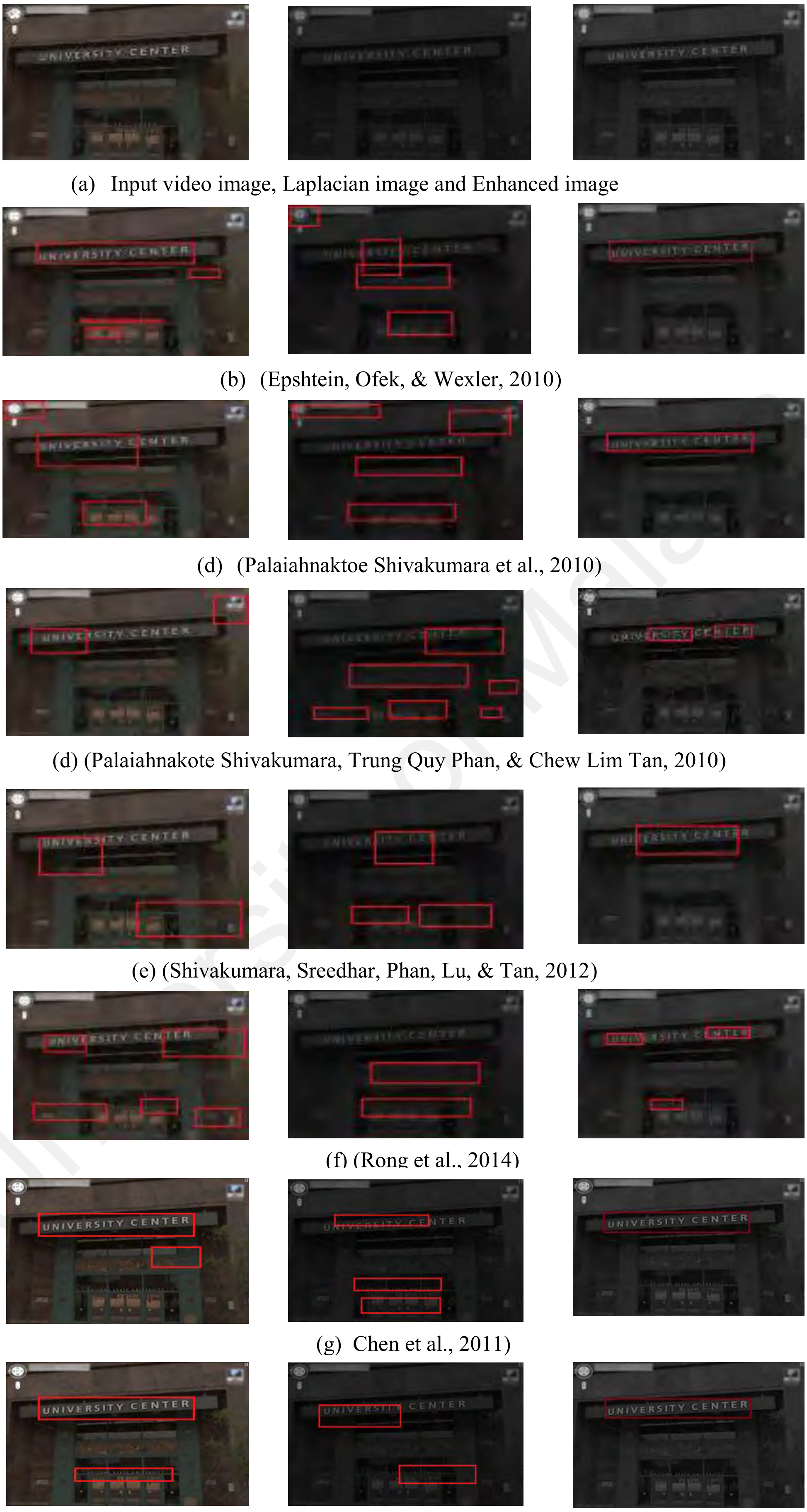
Figure 4.5: Sample qualitative results of the different text detection methods on input, Laplacian and enhanced images for ICDAR 2013 scene images.

Table 4.4: Text detection before and after enhancement on ICDAR 2013 Scene.

Methods	Before Enhancement						After Enhancement		
	Original			Laplacian			Enhanced		
	R	P	F	R	P	F	R	P	F
(Epshtein et al., 2010)	63.0	70.0	66.31	87.5	46.6	60.81	65.03	72.23	68.44
(Palaiahnaktoe Shivakumara et al., 2010)	49.0	58.0	53.12	49.5	40.4	44.48	60.21	54.12	57.00
(Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010)	61.37	41.51	49.52	50.27	36.51	42.29	74.23	41.24	53.02
(Shivakumara, Sreedhar, Phan, Lu, & Tan, 2012)	47.46	29.63	36.48	48.8	16.41	24.56	45.11	33.21	38.25
(Rong et al., 2014)	47.30	28.0	35.17	40.3	12.0	18.43	47.56	36.0	40.98
(H. Chen et al., 2011)	62.26	69.83	65.82	67.78	48.37	56.45	60.57	80.15	69
(Yin et al., 2014)	64.47	77.56	70.41	60.32	68.76	64.26	67.32	78.43	72.45

Table 4.5: Text detection before and after enhancement on SVT.

Methods	Before Enhancement						After Enhancement		
	Original			Laplacian			Enhanced		
	R	P	F	R	P	F	R	P	F
(Epshtein et al., 2010)	32.88	48.52	39.19	38.21	21.63	27.62	34.12	64.12	44.53
(Palaiahnaktoe Shivakumara et al., 2010)	48.1	30.41	37.26	50.6	20.26	28.93	50.47	39.12	44.07
(Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010)	28.12	15.2	19.73	18.37	10.34	13.23	32.18	22.17	26.25
(Shivakumara, Sreedhar, Phan, Lu, & Tan, 2012)	28.37	20.61	23.87	33.17	13.48	19.16	29.12	21.13	24.48
(Rong et al., 2014)	32.17	23.46	27.13	47.12	14.6	22.29	47.13	18.5	26.57
(H. Chen et al., 2011)	34.15	50.82	40.84	30.7	35.17	32.78	38.35	56.75	45.76
(Yin et al., 2014)	38.11	50.74	43.52	33.78	47.2	39.37	39.89	59.43	47.73



(a) Input video image, Laplacian image and Enhanced image

(b) (Epshtein, Ofek, & Wexler, 2010)

(d) (Palaiahnaktoe Shivakumara et al., 2010)

(d) (Palaiahnakote Shivakumara, Trung Quy Phan, & Chew Lim Tan, 2010)

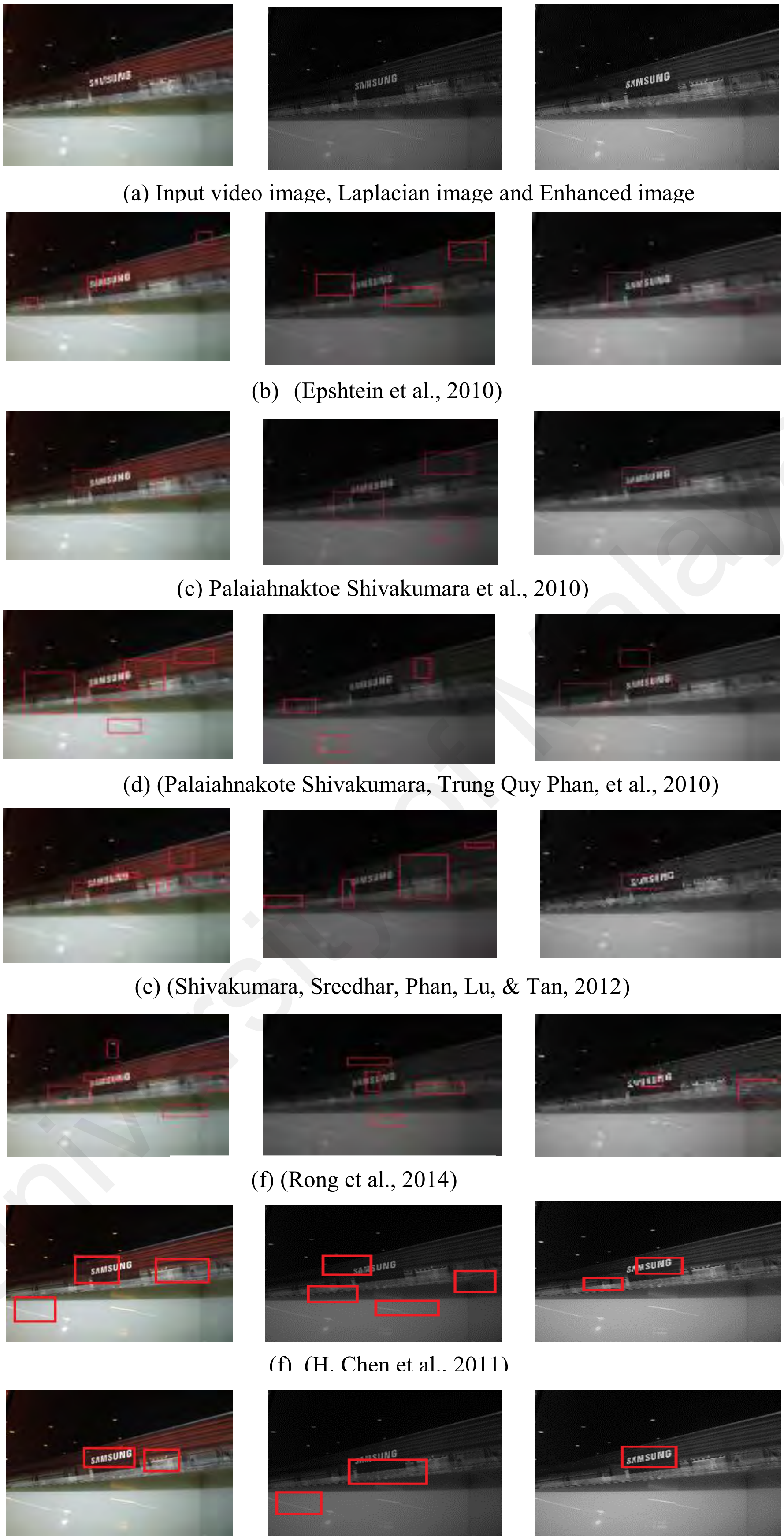
(e) (Shivakumara, Sreedhar, Phan, Lu, & Tan, 2012)

(f) (Rong et al., 2014)

(g) Chen et al., 2011)

(f) Yin et al., 2014)

Figure 4.6: Sample qualitative results of the different text detection methods on input, Laplacian and enhanced images for Street View Data.



(a) Input video image, Laplacian image and Enhanced image

(b) (Epshtein et al., 2010)

(c) Palaiahnaktoe Shivakumara et al., 2010)

(d) (Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010)

(e) (Shivakumara, Sreedhar, Phan, Lu, & Tan, 2012)

(f) (Rong et al., 2014)

(f) (H. Chen et al., 2011)

(f) Yin et al., 2014)

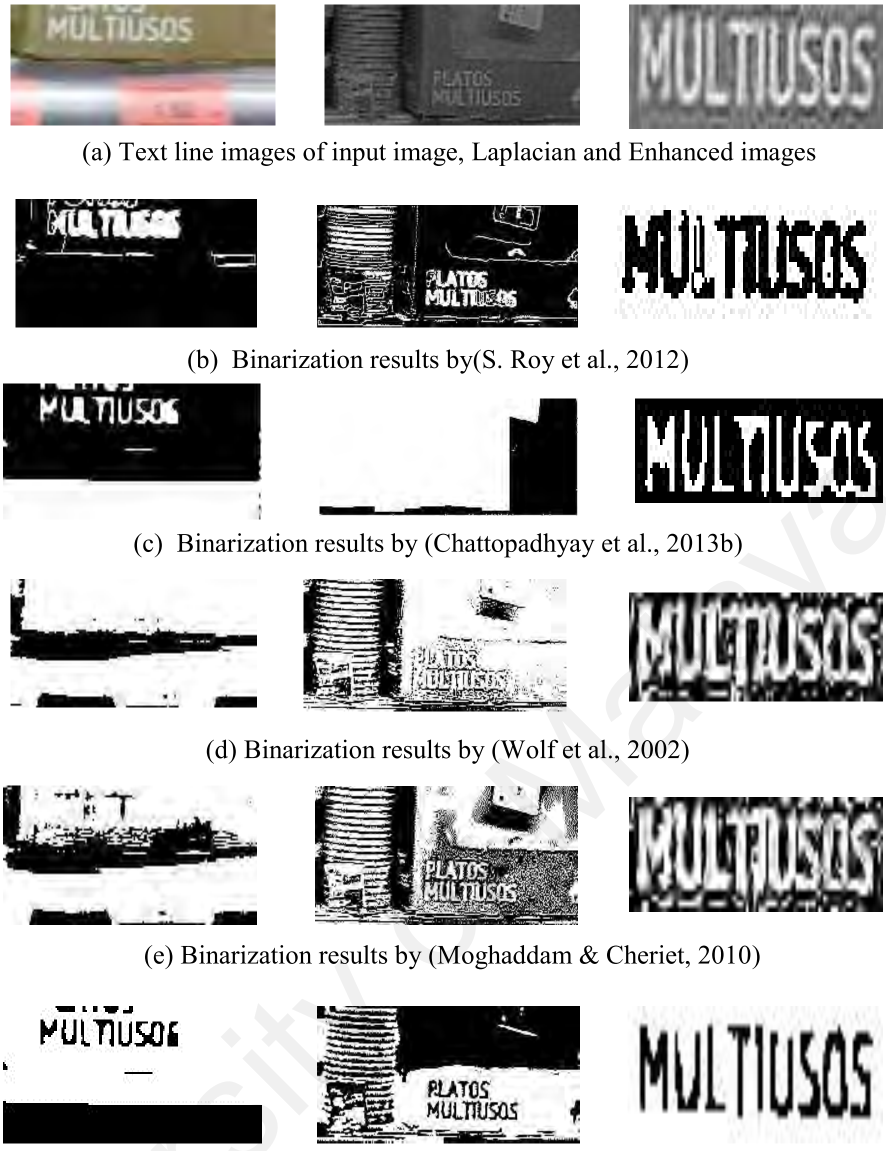
Figure 4.7: Sample qualitative results of the different text detection methods on input, Laplacian and enhanced images for MSRA data.

Table 4.6: Text detection before and after enhancement on MSRA.

Methods	Before Enhancement						After Enhancement		
	Original			Laplacian			Enhanced		
	R	P	F	R	P	F	R	P	F
(Epshtein et al., 2010)	30.65	24.12	26.99	31.2	15.35	20.57	32.5	24.12	27.68
(Palaiahnaktoe Shivakumara et al., 2010)	50.12	40.3	44.67	51.12	30.12	37.90	45.32	52.12	48.48
(Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010)	35.85	15.64	21.77	30.51	12.48	17.71	34.17	21.25	26.20
(Shivakumara, Sreedhar, Phan, Lu, & Tan, 2012)	47.72	31.1	37.65	21.13	27.5	30.22	50.14	38.7	43.68
(Rong et al., 2014)	41.43	17.0	24.10	28.17	17.45	21.55	36.13	28.12	31.62
(H. Chen et al., 2011)	36.48	29.73	32.76	28.39	24.47	26.28	35.26	34.57	34.91
(Yin et al., 2014)	57.38	63.7	60.37	53.65	55.58	54.59	59.18	64.73	61.83

In a similar way with text detection experiments, this section provides recognition experiments for the extracted text lines obtained by the detection techniques on the input, Laplacian, and enhanced images. The qualitative results of different binarization methods for different datasets, namely, ICDAR 2013 video, ICDAR 2013 scene, SVT, and MSRA, are shown in Figure 4.8 to Figure 4.11, respectively. It is noticed from Figure 4.8 to Figure 4.11 that the binarization techniques generate worst recognition even for the enhanced images on the ICDAR 2013 video and ICDAR scene datasets compared to those on SVT and MSRA. When the results of ICDAR 2013 video and ICDAR scene datasets are compared, the binarization methods give poor results on video data. This is because video data generally has low resolution compared to camera-based images as in ICDAR scene, SVT, and MSRA. Similarly, when the result of SVT and MSRA data is compared, the methods give poor results for MSRA but good results for SVT because MSRA contains arbitrarily oriented texts while SVT data contains almost horizontal texts. Sometimes, though binarization methods give good results for different oriented texts, OCR engine fails to recognize them correctly due to its inherent limitations. In addition, all the recognition approaches generate poor accuracy for the Laplacian images than the input and the enhanced images. This observation can be seen for all the dataset. When the recognition results of the enhanced and the input images are compared, the recognition rates of the enhanced images are better than those of the input images. The same conclusion can be justified by the quantitative results presented in Table 4.7 for all the four datasets. Table 4.7 depicts that the OCR results for Laplacian are lower than the input text line images, and the results of the enhanced images are higher than those of the input images. This shows that the presented model is efficient for increasing recognition accuracies.

In summary, from the above discussions, one can infer that the text detection and recognition performance degrade for the Laplacian images, at the same time, their performances are improved for the enhanced images significantly.



(a) Text line images of input image, Laplacian and Enhanced images

(b) Binarization results by(S. Roy et al., 2012)

(c) Binarization results by (Chattopadhyay et al., 2013b)

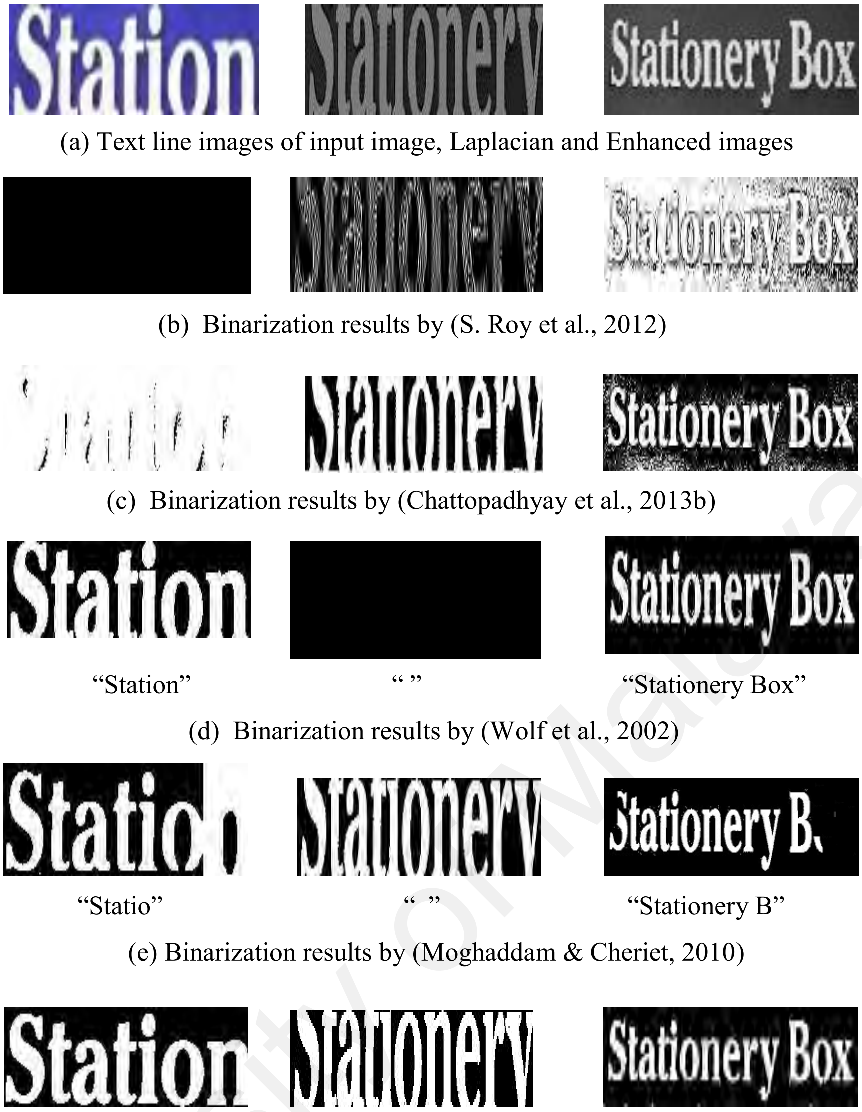
(d) Binarization results by (Wolf et al., 2002)

(e) Binarization results by (Moghaddam & Cheriet, 2010)

"MULTIUSOS"

(f) Binarization results by (Howe, 2013)

Figure 4.8: Recognition results of the binarization methods for the ICDAR 2013 video text line images. Note: since OCR engine gives nothing for the binarization results, recognition results are not reported for most of the texts except for the Howe's method.



(a) Text line images of input image, Laplacian and Enhanced images

(b) Binarization results by (S. Roy et al., 2012)

(c) Binarization results by (Chattopadhyay et al., 2013b)

"Station" " " "Stationery Box"

(d) Binarization results by (Wolf et al., 2002)

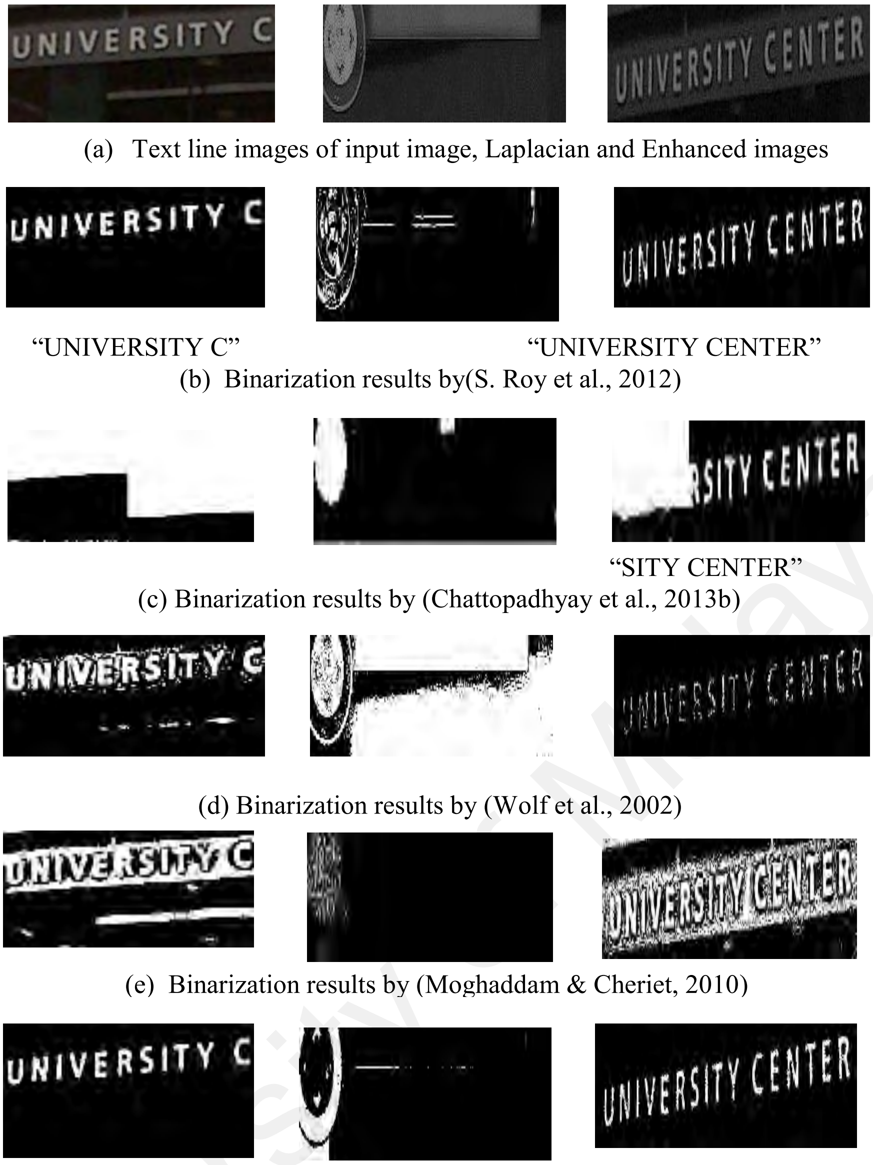
"Statio" " " "Stationery B"

(e) Binarization results by (Moghaddam & Cheriet, 2010)

"Station" " " "Stationery Box"

(f) Binarization results by (Howe, 2013)

Figure 4.9: Recognition results of the binarization methods for the ICDAR 2013 scene text line images. " " denotes recognition results by OCR engine nothing.



(a) Text line images of input image, Laplacian and Enhanced images

"UNIVERSITY C" "UNIVERSITY CENTER"

(b) Binarization results by(S. Roy et al., 2012)

"SITY CENTER"

(c) Binarization results by (Chattopadhyay et al., 2013b)

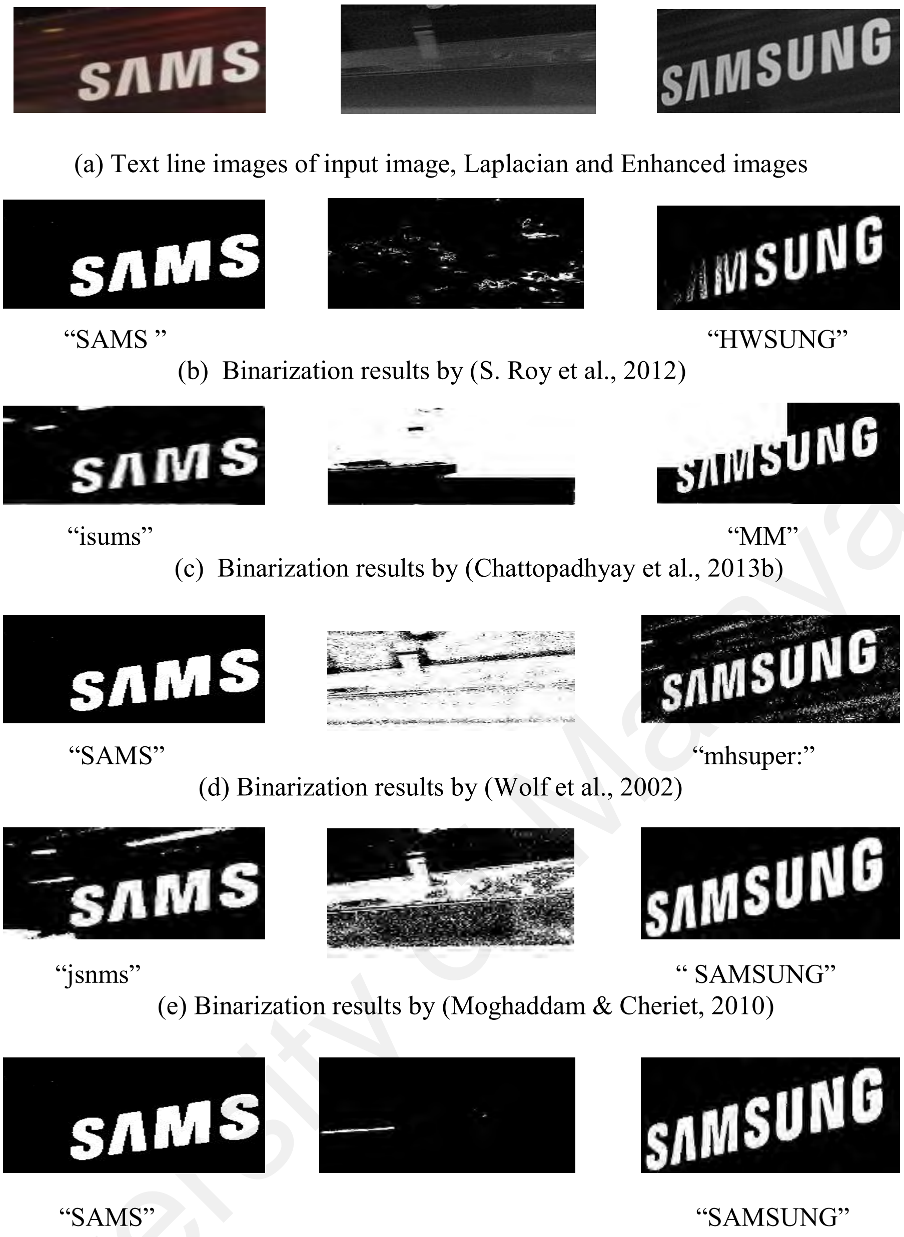
(d) Binarization results by (Wolf et al., 2002)

(e) Binarization results by (Moghaddam & Cheriet, 2010)

"UNIVERSITY C" "UNIVERSITY CENTER"

(f) Binarization results by (Howe, 2013)

Figure 4.10: Recognition results of the binarization methods for the SVT text line images. " " denotes recognition results by OCR engine nothing.



(a) Text line images of input image, Laplacian and Enhanced images

“SAMS ” “HWSUNG”

(b) Binarization results by (S. Roy et al., 2012)

“isums” “MM”

(c) Binarization results by (Chattopadhyay et al., 2013b)

“SAMS” “mhsuper:”

(d) Binarization results by (Wolf et al., 2002)

“jsnms” “ SAMSUNG”

(e) Binarization results by (Moghaddam & Cheriet, 2010)

“SAMS” “SAMSUNG”

(f) Binarization results by (Howe, 2013)

Figure 4.11: Recognition results of the binarization methods for the MSRA text line images.

(Note: Since OCR engine returns null for the binarization results, there are no recognition results for this dataset.)

Table 4.7: Recognition results Before Enhancement (BE) and After Enhancement (AE).

(Note: Shivakumara et al-wavelet (Palaiahnaktoe Shivakumara et al., 2010) is used for text line detection from both original, Laplacian and denoising image. Here RR: Recognition Rate, O: Original Image, L: Laplacian Image, E: Enhanced Image)

Methods	ICDAR 2013 Video			ICDAR 2013 Scene			SVT			MSRA		
	BE		AE	BE		AE	BE		AE	BE		AE
	O	L	E	O	L	E	O	L	E	O	L	E
	RR	RR	RR	RR	RR	RR	RR	RR	RR	RR	RR	RR
(S. Roy et al., 2012)	21.12	16.12	35.7	24.12	21.6	48.12	21.32	14.12	37.1	16.5	11.7	18.12
(Chattopadhyay et al., 2013b)	23.68	23.8	36.12	21.68	22.68	47.76	23.18	12.21	34.21	17.1	10.8	19.74
(Wolf et al., 2002)	24.1	18.1	32.51	25.1	19.1	42.83	23.6	15.12	36.24	16.5	11.4	18.8
(Moghaddam & Cheriet, 2010)	27.1	24.7	39.34	22.1	20.52	50.12	22.5	18.27	41.1	15.12	11.2	17.23
(Howe, 2013)	12.2	26.5	41.2	29.8	23.2	53.35	27.2	17.57	43.12	18.6	14.5	21.41

4.5 Summary

In summary, a new Fractional Poisson model has been proposed for increasing fine details in natural scene images as well as video by considering Laplacian noise. As the presented enhancement model gives a satisfactory result for the Laplacian noise, it also performs well for manually added noise and distortion. The proposed model yields better quality measure than those existing standard techniques for all the four benchmark datasets. To validate the efficacy of the presented model, different text detection and recognition methods are applied on the input, Laplacian, and enhanced images. The quantitative and qualitative results confirm that the presented approach aids in enhancing the performances of text extraction and recognition, significantly.

CHAPTER 5: TEXT DETECTION AND KEYWORD SPOTTING IN VIDEO

5.1 Background

The previous chapter presents a method for increasing the contrast of video text pixels based on Fractional Poisson model. This chapter focuses on detecting text and spotting keyword in Marathon, race, and sports. Therefore, a multi-modal technique that fuses both biometric and textual features has been proposed for identifying bib number/text. The words, detected using text detection approach are used for spotting the keyword in the video. A novel approach based on texture-spatial feature and context feature have been proposed for keyword spotting in the scene, video, and license plate images to retrieve desired information without recognizing text.

5.2 Word/Bib Number Detection in Video Images

A multi-modal technique that combines face, skin and text detections has been proposed for bib number to achieve better results in detection and recognition. This is valid because a person who participates in Marathon and Olympic sports generally displays skin at the leg, face, neck, etc. With this cue, person body which can cover bib numbers can be identified without missing and irrespective of body turn. This step results in text candidate regions. The above step eliminates most backgrounds and gives the regions that contain bib numbers/text. However, the above step fails to identify the bib numbers/texts accurately since backgrounds also come with the text. In this way, this step helps text detection and recognition techniques to detect or recognize bib numbers/texts accurately by removing complex background successfully. Therefore, a bib number/text detection method has been proposed from the candidate text regions. Then detected text lines are passed to OCR to get recognition results through binarization technique. The details of approach have been discussed in following sub-sections.

5.2.1 Text Candidate Region Detection

It is true from skin detection that color plays a vital role in identifying skin color in images (Conaire, O'Connor, & Smeaton, 2007; Kakumanu, Makrogiannis, & Bourbakis, 2007). Since inputs are the marathon and running race images, definitely the skin of persons appears in the images. Therefore, skin color detection is explored to identify person bodies in the input images. The main advantage of skin color detection is that it does not depend on faces as the existing technique does (Ben-Ami et al., 2012) for text candidate region detection. For the given image, the proposed technique detects faces if faces are present in the image using Open CV implementation (Lienhart & Maydt, 2002). This results in rectangular boxes for the faces in the image. For each pixel in the rectangular box, the proposed technique computes the mean for R, G and B values corresponding to the pixels in each rectangular box. The means of respective color values is considered as a seed, which is represented by the following equation (5.1),

$$Seed = (\ R_m,\ G_m,\ B_m\ ) \qquad (5.1),$$

where R_m, G_m, and B_m are respective seed values for the R, G, B pixels in the rectangular box.

Since these seed color values are extracted from a facial region, it can be believed that the seed values represent the skin color of the person in the image. The proposed technique compares the values of the pixels in the input image with the seed color values and classifies them into respective seed value clusters when a pixel value is close to a respective seed color value. This results in three clusters as defined in equation (5.2)-equation (5.4).

$$C_R = \{\ P\ |P.R\ \in [\ Seed.R_m - \Delta, Seed.R_m + \Delta]\ \} \qquad (5.2)$$

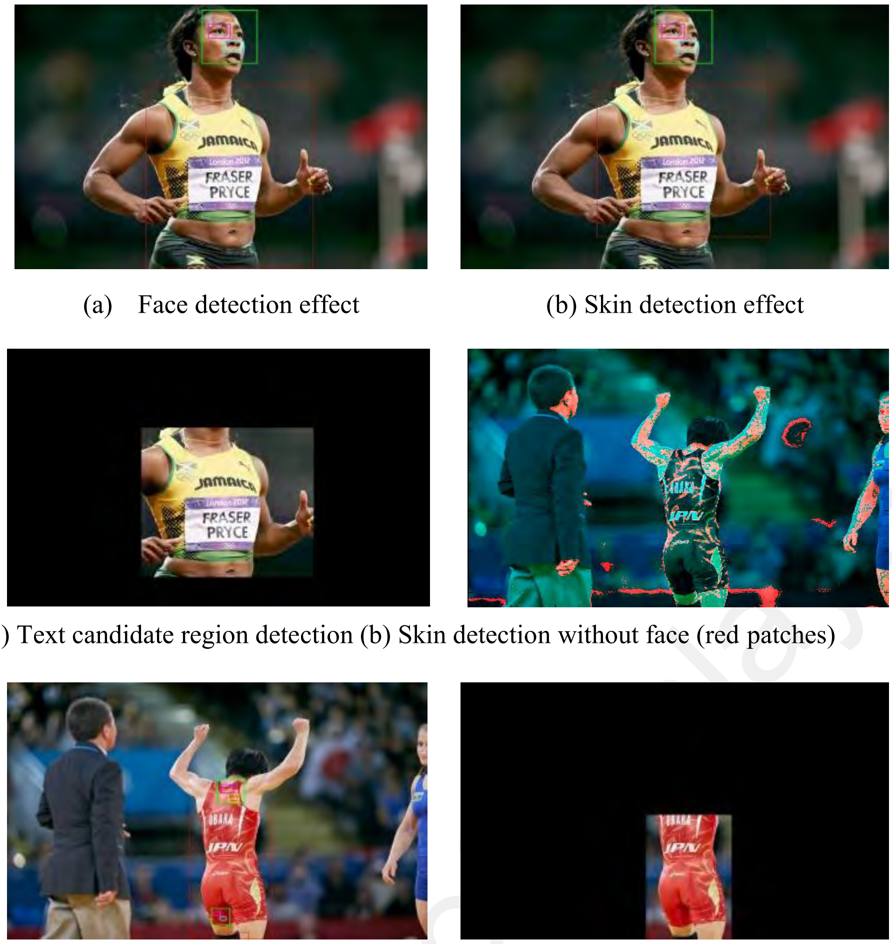
$$C_G = \{\ P\ |P.G\ \in [\ Seed.G_m - \Delta, Seed.G_m + \Delta]\ \} \qquad (5.3)$$

$$C_B = \{\ P\ |P.B\ \in [\ Seed.B_m - \Delta, Seed.B_m + \Delta]\ \} \qquad (5.4)$$

Where $C_R, C_G, and\ C_B$ denote three color clusters, P is a pixel in the input image, and Δ is a threshold which is set to the minimum distance with the particular cluster. The proposed technique combines all the three cluster values as one cluster to obtain the regions that represent skin of the person in the image. Thus the combined cluster can be defined in equation (5.5).

$$C_C = C_R \cup C_G \cup C_B \tag{5.5}$$

where $\boldsymbol{C_C}$ represents the combined cluster. Suppose, the face does not present in the image, the existing skin detection method (Conaire et al., 2007) has been used for identifying skin candidates and then employ the above procedure to detect skin regions of persons in the image. The skin detection method (Conaire et al., 2007) adapted in this work is based on a non-parametric histogram model. This model is trained using supervised learning from the skin and non-skin pixels. Since the detected region covers the whole body of the person including legs, definitely, the region includes texts or bib number tags which may appear at any part of the body. However, when the image contains several persons, the skin regions detected by the above procedure may be merged. In order to obtain skin area which covers texts of a person, the same hypothesis has been proposed as used in (Ben-Ami et al., 2012) for torso detection using skin candidate regions, which results in text candidate regions. This procedure is illustrated in Figure 5.1, where (a) shows the detection result of face and its torso, (b) depicts the outcome of face and skin detection, (c) represents the final result of the detected text candidate region, (d) is the result of skin detection for the image where there is no face, (e) is result of torso detection using skin candidates, and (f) is the final text candidate region for the image that has no faces. In this way, the proposed technique can detect text candidate regions from the image either with a face or without a face.



(a) Face detection effect (b) Skin detection effect

(c) Text candidate region detection (b) Skin detection without face (red patches)

(e) Torso detection using skin candidate (f) Text candidate region detection

Figure 5.1: Text candidate region detection using both face and skin.

5.2.2 Multimodal based Approach for Word/Bib number detection

Figure 5.1 shows text candidate regions, which contain text or bib number. It can be noticed from the results that though a text candidate region contains text or bib number, the region covers lots of background information, which cannot be used for recognition. Therefore, a text detection method is proposed to locate the texts in the candidate regions. Since this section considers text candidate region detected by the biometric features for text detection, it is named as a multi-modal method. Since there are plenty of text detection techniques available in the literature, the technique (Palaiahnaktoe Shivakumara et al., 2010) which works well for the marathon and running race images has been used. The fusion of color, wavelet, and k-means clustering has been used in (Palaiahnaktoe

Shivakumara et al., 2010). Since the technique uses the color feature as the main attribute for identifying probable text candidates, it is believed that the same color plays a vital role in detecting text in the sports and running race frame. In addition, the technique works well for the situations such as low contrast, complex background, low resolution and arbitrary orientation. These characteristics of the text detection technique motivate to propose the technique for detecting bib numbers or texts in the text candidate regions in this work.

In the same way, there are lots of techniques for binarizing scanned, camera-based and video images in the literature. Since the primary focus is to show how the combination of biometric and textual features help in improving detection and recognition of text in the sports images compared to text features alone, the existing techniques which are suit for this work are preferred to use. With this notion, the binarization technique in (Howe, 2013), which tunes the parameters automatically for binarizing images has been proposed to use. It is known that the main issue of a binarization technique is to control the threshold and parameters used when different data or applications are given as the input. Since it is hard to define parameters or threshold values for the current bib detection in sports images, this binarization method has been used for binarizing bib numbers or texts.

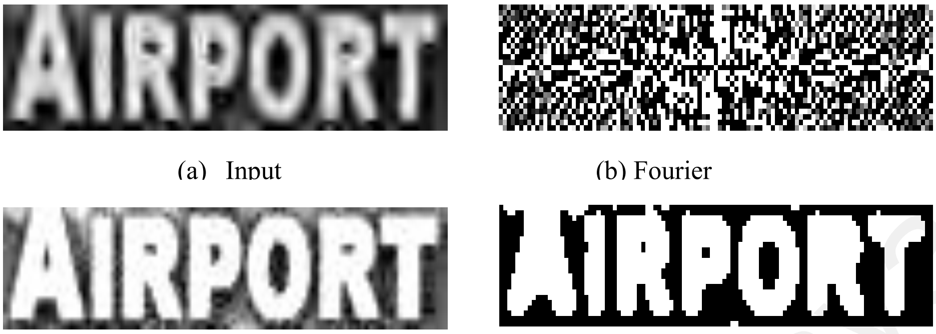
In overall, a new multimodal technique that combines both biometric and textual features has been proposed for detecting text to enhance the recognition. Face and skin features are explored in a new way of identifying text candidate regions for input images. However, text retrieval approach through the biometric feature, text detection, and text recognition might not be sufficient to handle different type’s text present in the natural scene, video and license plate. Sometimes, recognition based approach fails to retrieve the semantic content of text where various style and size of the font, and contrast, etc. are observed. Therefore, to accelerate the retrieval process without text recognition, spotting based approach has been proposed in the following section.

5.3 Word Spotting in Images

Keyword spotting in video document images is difficult due to complex nature of video which is discussed in Introduction Section. For the segmented words, a set of texture features has been proposed to extract the global textual properties. Since texture represents global information, it is invariant to the arbitrary orientation of words and font or font size variations. The set of texture features are used for identifying text candidates of word images using k-means. Since texture extracts the global pattern of a word image, the same features cannot be used for spotting words in the video. Therefore, new features have been proposed based on spatial arrangements of pixels in text candidate images to spot the words in the video.

5.3.1 Texture Features for Text Candidate Selection

Since a video word image suffers from low resolution and low contrast, enhancement has been carried out on low contrast information in the word image by obtaining the Fourier image for the input image. Then Gaussian filtering is applied to remove noisy frequency coefficients introduced by Fourier process. The reason to use Fourier transform is that Fourier generates high-frequency coefficient in case of high contrast and low-frequency coefficient for low contrast. This results in the enhancement of text pixels in the word image as shown in Figure 5.2. Here input word is presented in (a), and the result of the Fourier spectra which gives horizontal and vertical bright directions is presented in (b), and the result of Gaussian filtering which is brighter than the input image is depicted in (c). In this way, Gaussian filtering over Fourier co-efficient enhances text pixels. To segregate non-text pixels from text ones, texture features has been proposed because it is true that texture features give high energy for high contrast and low energy for low contrast pixels. Therefore, 10 features, namely, energy, entropy, inertia, local homogeneity, mean, the second-order and the third-order central moments, and median, the second-order and the third-order median moments are computed as given respectively from equation (5.6) to equation (5.15). A sliding square window is moved over the enhanced image. The size of window is 3×3. The above features are reckoned from each window as follows:



(a) Input

(b) Fourier

(c) After filtering

(d) Text candidates

Figure 5.2: Text candidate selection using texture features.

$$E = \sum_{i,j} W^2(i,j) \tag{5.6}$$

$$Et = \sum_{i,j} W(i,j).logW(i,j) \tag{5.7}$$

$$I = \sum_{i,j} (i-j)^2 W(i,j) \tag{5.8}$$

$$Hm = \sum_{i,j} \frac{1}{1+(i-j)^2} W(i,j) \tag{ 5.9}$$

$$M = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} W(i,j) \tag{5.10}$$

$$\mu_2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (W(i,j) - M)^2 \tag{5.11}$$

$$\mu_3 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(W(i,j) - M)^3 \tag{5.12}$$

$$M\mu = SW(\frac{N^2+1}{2}) \tag{5.13}$$

$$Me_2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(W(i,j) - M\mu)^2 \tag{5.14}$$

$$Me_3 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(W(i,j) - M\mu)^3 \tag{5.15}$$

Here, at pixel position (i,j), the sliding window W of size N×N is placed and SW is the sorted pixel values of W. After feature computation, $10 \times \text{row} \times \text{column}$ features are obtained, which are called as a feature vector. The k-means algorithm with k = 2 has been applied on feature vectors to make two cluster groups: text candidates and background. The cluster with high average is considered as text candidates as displayed in Figure 5.2 (d), where a clear text portion (white parts) is seen for the word displayed in Figure 5.2 (a).

5.3.2 Texture-Spatial Feature for Word Spotting in Images

Since texture property represents the global appearance pattern of text in the word image, the step presented in the previous section may not distinguish different words of the same font, font size, etc. Therefore, it is needed to study the local information of the obtained text candidates. Hence, the text candidate image is divided into four parts, denoted as P1-P4, in a new way as shown in Figure 5.3 (a). The reason to divide the text candidates in this fashion is that the distribution of text pixels is generally rich in horizontal, vertical and diagonal directions. This type of division adds a more discriminative power to the features. The actual segmented parts (P1-P4) for the text candidate image are displayed in Figure 5.3 from (b) to (e), respectively. For each part, a proximity matrix is estimated as in equation (5.16) using distance calculated among pixels to study the spatial relationship between the pixels because the spatial arrangements make a difference among different words. The mean and the standard deviation are computed as in equation (5.17) and equation (5.18) for each normalized proximity matrix of the parts, which results in a feature vector having 8 sub-features (4 means and 4 standard deviations) for each word. The feature vector of the query word is compared with the feature vector of every word in the database to find matched words using cosine symmetry as defined in equation (5.19). As it is suggested in (Kesidis & Gatos, 2011) that the cosine metric is useful for keyword spotting, the same distance metric as a similarity measure has been proposed for spotting the words in this work. The mathematical details for proximity matrices estimation and features extraction for each part is given below.

The proximity matrix of part P1 is represented by $Part1_{(i,j)}$ and is estimated as in equation (5.16), where P is the set of all pixels in the P1 portion and P' is the transpose of P:

$$Part1_{(i,j)} = \sum_{r=1}^{n} \sqrt[2]{{P_{i,r}}^2 + {P'_{r,j}}^2} \tag{5.16}$$

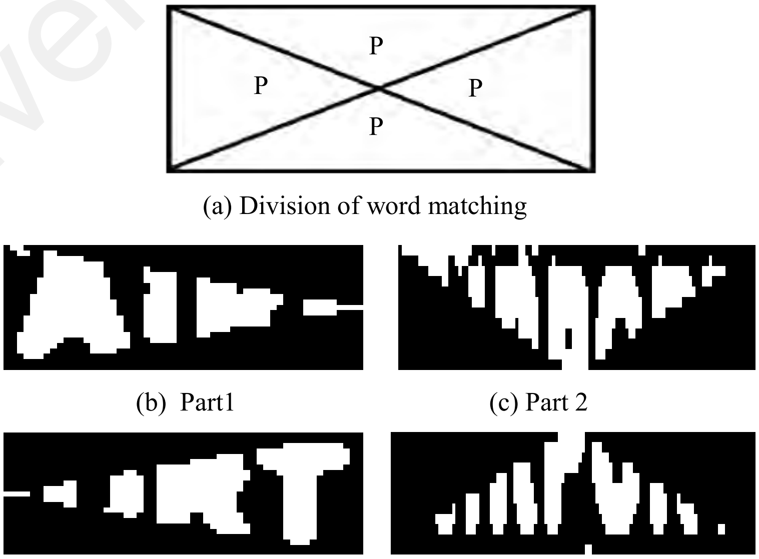
The proximity matrices of P2, P3, and P4 are represented by $Part2_{(i,j)}$, $Part3_{(i,j)}$ and $Part4_{(i,j)}$, respectively, and are calculated following equation (5.16). Similarly, the mean and the standard deviation for proximity matrices are computed. The mean of P1 portion pixel distances is defined by $Mean(Part1)$, and the standard deviation of P1 portion pixel distances is represented by $Std(Part1)$ where Part1 is the set of all pixel distances of P1 portion and n is the number of the distances in Part1.

$$Mean(Part1) = \frac{1}{n}\sum_{i=1}^{n} Part1_i \tag{5.17}$$

$$Std(Part1) = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}(Part1_i - \ Mean(Part1))^2} \qquad (5.18)$$

The mean of the pixel distances of P2, P3, and P4 portion distances are represented by $Mean(Part2),\ Mean(Part3)$ and $Mean(Part4)$, respectively, and are computed as in equation (5.17). The standard deviations of P2, P3, and P4 portion pixel distances are represented by $Std(Part2),\ Std(Part3)$ and $Std(Part4)$, respectively, and computed as in equation (5.18). The cosine distance metric is computed for the query word as well as database words as follows. Every segmented word is represented by a feature vectorp_i, $1 \leq\ \mathrm{i} \leq \mathrm{n}$ with k features, while the input query is represented by a feature vector q with k features. Then the distance between p_i and q is measured as in equation (5.19), where p_{ij} and q_j are the j-th features of p_i and q, respectively. Words with smaller distance value are more similar to the query.

$$Dist_i = 1 - \frac{\sum_{j=1}^{k} p_{i\,j}\ q_j}{\sqrt{\sum_{j=1}^{k} p_{i\,j}^2 \sum_{j=1}^{k} q_j^2}} \qquad (5.19)$$



(a) Division of word matching

(b) Part1

(c) Part 2

(d) Part 3

(e) Part 4

Figure 5.3: Local information extraction for matching.

However, the approach is not robust to the images affected by modern urban environments such as posters, product tags, license plates, electronic signs, guideposts, and billboards, which often contain texts in different forms and background, therefore, only a combination of texture-spatial feature might not be sufficient to handle variety of text. Therefore, a novel approach which introduces fractional means based context features has been proposed for spotting keywords in video frames, natural scene images and license plate images.

5.4 Word Spotting in Video

In this section, a novel method integrating fractional means along with context feature namely, Radon and Fourier coefficients have been proposed for word spotting in different types of data.

Since word extraction from video, natural scene, and license plate images is an important component for word spotting, a new method has been proposed for extracting words from text lines. In this work, the fractional means feature has been explored for the image and then use k-means clustering with k=2 for segregating non-text candidates from text candidates because it is a fact that appearance of character shapes is considered as special texture property (Shivakumara et al. 2015). As a result, fractional means generates low values and high values for non-text pixels and text pixels, respectively. For this reason, k-means clustering has been used for classification. However, due to variations in low resolution and background in video, natural scene, and license plate images, it is hard to segregate non-text candidates from text candidates by fractional means feature alone, and hence it misclassifies non-text as text. As noted in the work proposed in (Retsinas et al. 2016) for keyword spotting based on gradient projections, where it is mentioned that for text pixels, the combination of Radon and Fourier coefficients in gradient domain helps in extracting regular pattern of text which represent unique geometrical feature of the text while it extracts non-regular pattern for non-text, which fails to represent text components. This observation is explored for defining context features by finding a relationship between foreground and background of text candidates given by the previous step, which eliminates most of the non-text candidates and hence the result is considered as text representatives. Due to unpredictable nature of video, natural and license plate images, there are chances of losing text information by the previous step. Therefore, a new step called minimum cost path is proposed based ring growing which restores the missing text information using the Canny edge, segments the words from the text lines and finally, and eliminates false positives based on the relationship between forward and backward path information of the text lines. Further, the proposed method considers the above mentioned Radon and Fourier coefficients as features, and it extracts for both foreground and background of the words for performing spotting in the images using distance measure.

5.4.1 Fractional Means Features for Detecting Text Candidates

The fractional means has been popular in mathematics as it converges faster to the population mean whenever this mean exists. As a result, fractional means preserves convergent sequences and their limits. If the sequence of the fractional means is convergent, then the series is said to be fractional summable which results in meaningful information about the population. The same property has been explored for studying the behavior of the character component in this work such that text candidates can be separated from non-text candidates. Therefore, mathematically, it can be formulated for the image is as follows. Here a sequence is defined as a collection of pixel values of a particular block (window) of the image. Let (A. Sharma) be a sequence of image pixels, and let

$$\rho_1 = c_1, \quad \rho_2 = c_1 + c_2, \ldots, \quad \rho_k = c_1 + \cdots + c_k$$

be the k-th partial sum of the series $S = \sum_{i=1}^{\infty} c_i$. The series S is called the fractional summable.

The fractional means is defined as follows:

$$\sigma_k = \frac{1}{k}\sum_{n=1}^{k}\rho_n$$

$$\sigma_k = \frac{1}{\mathrm{k}}[\rho_1 + \cdots + \rho_{\mathrm{k}}]$$

$$\sigma_{\mathrm{k}} = \frac{1}{\mathrm{k}}[\mathrm{c}_1 + (\mathrm{c}_1 + \mathrm{c}_2) + \cdots + (\mathrm{c}_1 + \cdots + \mathrm{c}_{\mathrm{k}})]$$

$$\sigma_{\mathrm{k}} = \frac{1}{\mathrm{k}}[\mathrm{kc}_1 + (\mathrm{k}-1)\mathrm{c}_2 + (\mathrm{k}-2)\mathrm{c}_3 + \cdots + \mathrm{c}_{\mathrm{k}}]$$

$$\sigma_k = \sum_{n=1}^{k}\frac{k+1-n}{k}\quad c_n \tag{5.20}$$

where $\sigma_{\mathrm{k}}$is called the fractional means, where k is the image block length, and ρ_{n} is the gray value of corresponding image pixels. In this way, fractional means is calculated for each sliding window over the image to extract features. However, the convergence property of fractional means can be proved as follows.

If c_n → pixel value (v) of corresponding image block for all n=1,…,k then σ_{k} is converged to v.

Proof. Let ε > 0. Since $c_n \rightarrow v$, there is a natural number N such that | c_n - v | < ε/2 for every index n ≥ N. Suppose that

$$\sigma_n = \frac{(c_1+\cdots+c_{N-1})}{n} + \frac{(c_N+\cdots+c_n)}{n} \tag{5.21}$$

Then,

$$|\sigma_n - v| = \left|\frac{((c_1 - v) + \cdots + (c_{N-1} - v))}{n} + \frac{((c_N - v) + \cdots + (c_n - v))}{n}\right|$$

$$\leq \frac{(|c_1 - v|) + \cdots + |c_{N-1} - v|)}{n} + \frac{(|c_N - v| + \cdots + |c_n - v|)}{n}$$

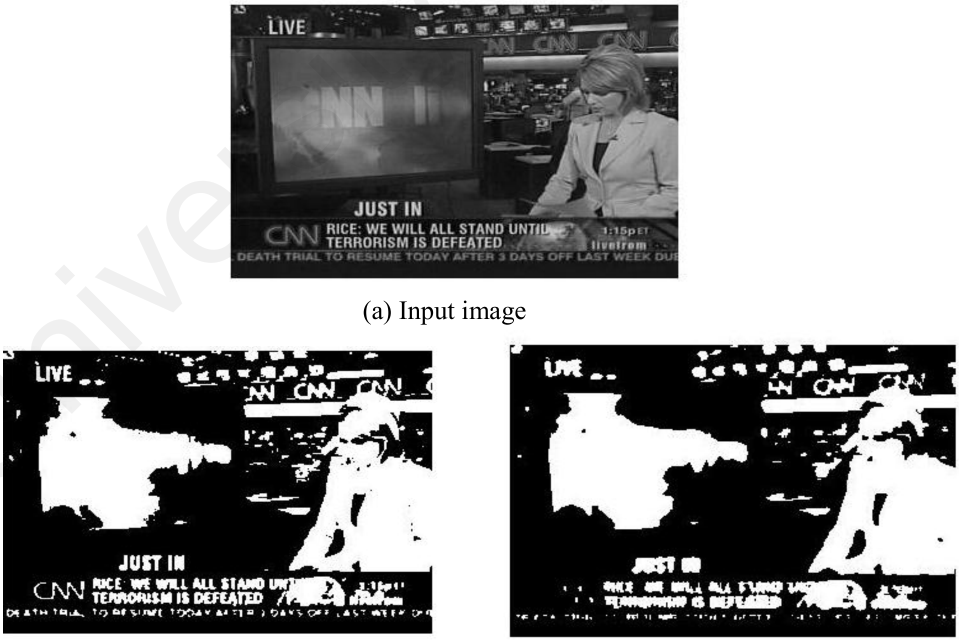
Since | c_n - v | < ε/2, for all n=1,…,k, thus it is obtained

$$|\sigma_n - v| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Hence, $\sigma_n \rightarrow v$, which completes the proof.

The above proof shows that the fractional means converges to the value of the pixel, whenever the sequence of the image pixels converges to its value. This is an advantage of fractional means.

The proposed method extracts fractional means features for each sliding window over the image which results in feature matrix of the same dimension of the input image. Since fractional means generates low and high values for segregating non-text pixels from text pixels, k-means is proposed to use with k=2 clustering which outputs two clusters. The cluster which yields high average than the average of other cluster is considered as a text. In this work, the block or window size is determined empirically as 3×3. One sample results for the different window is shown in Figure 5.4. The input image is displayed in Figure 5.4 (a). Text cluster result given by the proposed method is good for a 3×3 window, shown in Figure 5.4 (b) compared to the results of a 5×5 window which lose significant information of the text as shown in Figure 5.4 (c).


LIVE
JUST IN
RICE: WE WILL ALL STAND UNTIL
TERRORISM IS DEFEATED
1:15pET
DEATH TRIAL TO RESUME TODAY AFTER 3 DAYS OFF LAST WEEK DUE

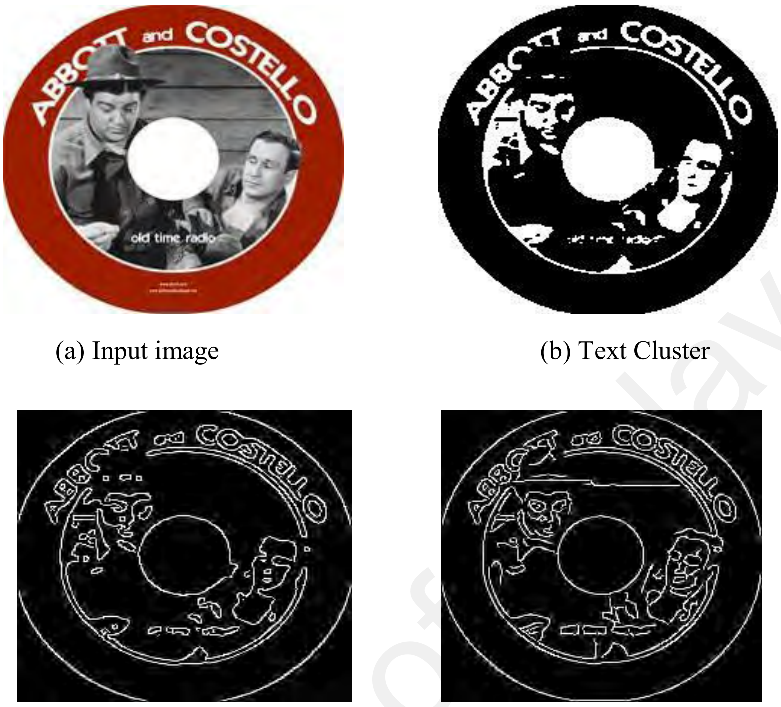
(a) Input image

(b) Text cluster for 3×3 window

(c) Text cluster for 5×5 window

Figure 5.4: Sample illustration for determining size of the window for extracting Cesaro means features.

The effect of k-means clustering on Ceasro means features with a 3×3 window for the input image can be seen in Figure 5.5, where for the input image containing curved texts with the complex background is shown in Figure 5.5 (a). The proposed approach obtains text cluster which is displayed in Figure 5.5 (b), where text pixels are identified correctly. Due to contrast variations from one image to another, k-means on Cesaro means sometimes represents background with black color, foreground with white color and vice versa. To overcome this problem, Canny edge image has been proposed of the text cluster given by the proposed approach as shown in Figure 5.5 (c), where it can be noticed that all the text pixels are represented by white pixels, while non-text pixels are represented by black pixels. This is true for the text cluster containing background represented by white color and foreground represented by black color. This is valid because Canny edge detector considers contrast information for finding edges. In case the proposed step misses the structures of character components due to poor contrast and noisy background, restoration is proposed for missing information by extracting edge components in the Canny edge image of the input image as shown in Figure 5.5 (d), where it can be seen that missing pixels are restored such that structures of character components will be preserved compared to Figure 5.5 (c), which are called text candidates. As discussed in the above regarding advantage of the Canny edge image of text cluster, one example is displayed in Figure 5.6. Here, a input image is shown in Figure 5.6 (a), and the text cluster given by the proposed approach contains background represented by white color and foreground represented by black color in contrast to the text cluster results shown in Figure 5.5 (b). However, when the Canny edge images of the text cluster results in Figure 5.5 (b) and Figure 5.6 (b) are compared, which are respectively depicted in Figure 5.5 (c) and Figure 5.6 (c), it can be confirmed that foreground pixels (text pixels) are represented by white color for both the text cluster results. In this work, the Canny edge detector is preferred than other edge detectors likes, Pewit, Sobel, and Laplacian, etc. for finding edge images because it is good for low contrast as well as high contrast images.



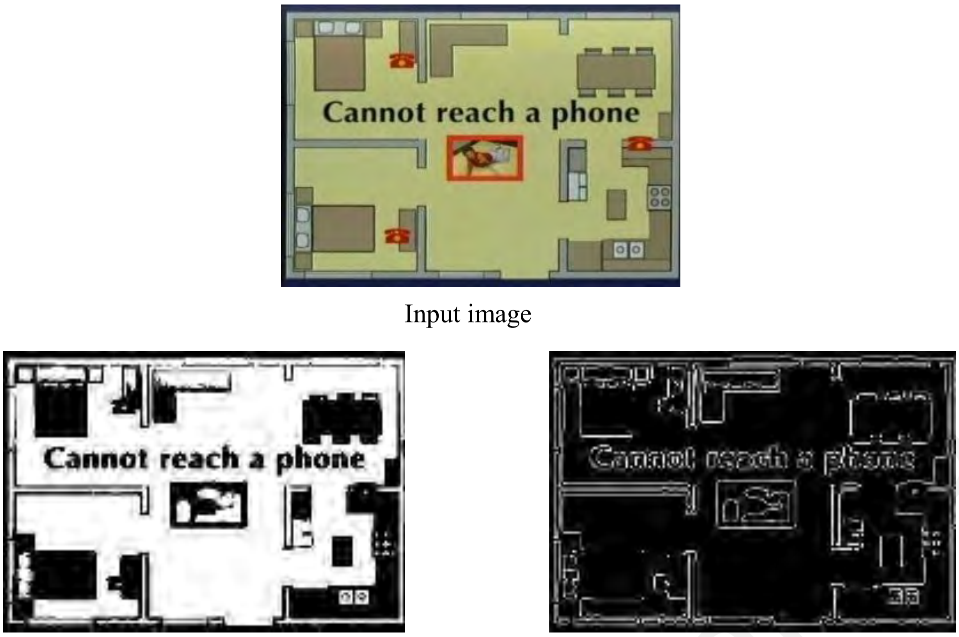
(a) Input image

(b) Text Cluster

(c) Canny of the (a)

(d) Text Candidates

Figure 5.5: Text candidate detection based on Cesaro means features with k-means clustering.



Input image

(b) Text Cluster Results

(c) Canny edge of (b)

Figure 5.6: Sample illustration of the Canny edge image for changing background and foreground colors.

5.4.2 Context Features for Detecting Text Representatives

It is observed from the text candidate described in the previous section that the above step identifies non-text candidates as a text due to unpredictable nature of input images. Therefore, context feature is proposed based on foreground and background of each text candidate to eliminate false text candidates in this step. Here the pixels which represent white color and the pixels which represent black color are considered as foreground and background, respectively. It is stated in (Retsinas et al., 2016) for spotting words in handwritten document images that dividing gradient into several directions can provide local information such as vertical, horizontal and diagonal, which helps to extract more details of the structures of character components. This observation motivates to divide the gradient image computed as defined in equation (5.22) of text candidates into 0, 45, 90 and 135 directional images as defined in equation (5.23), and the pixels in the gradient image that gives respective angles are displayed as white pixels in the respective directional images as displayed in Figure 5.7 (a) - (b). In other words, the proposed approach divides the gradient image of text candidates into directional images, which results in respective directional binary images. To extract the structure of text candidates, Radon transform has been proposed to use as defined in equation (5.24) for selected angles, namely, 0, 30, 60, 90, 120 and 150, which are denoted as projection angles on each directional image including the source gradient image as suggested in (Retsinas et al., 2016) for keyword spotting in handwritten document images. Since the considered problem here is much more complex, Radon transform may introduce noise coefficients. To smooth such coefficients, Fourier transform has been applied to obtain Fourier coefficients as it filters low-frequency coefficients that represent noise pixels in text candidate images. It is true that because of complex background with high contrast, sometimes, Fourier also gives high coefficients for non-text pixels as text ones. It is believed that gradient values of edge pixels of text candidates usually have the same values because of the fact that the pixels of the whole character component have uniform color values. With this observation, to eliminate such coefficients, line graphs are plotted for the coefficients of each projection angle of each directional image and the source gradient image. In Figure 5.7 (c), one sample illustration is given. Here, the line graph is drawn for Fourier coefficients on the Radon projections for source gradient image. The proposed approach performs clustering, which groups the coefficients that represent the same length of peaks with a certain threshold as shown in Figure 5.7 (c), where the same length peaks are marked by red color. In other words, the first peak is compared with all the other peaks to select the peaks which have the same length the first peak with a certain threshold. This results in one cluster. Then the second peak with all the other peaks results in the second cluster and so on. Further, the proposed approach chooses Fourier coefficients of the cluster that gives the minimum standard deviation among several clusters as features for the projection angle of the directional image as shown in Figure

5.7 (d), where cluster 2 gives the lowest standard deviation among several clusters. This is valid because the peaks which represent edges of text candidates must have the same value and the peaks which represent non-text pixels have variations in coefficient values. In this way, the proposed approach extracts feature coefficients for projection angles of all the directional images. The concatenation of the feature coefficients of all the directional images is considered as the feature vector for text candidates. When the directional images are observed shown in Figure 5.7 (b), the union of the information in directional images gives back to the original image. Therefore, the feature coefficients extracted from those directional images are considered as local features.

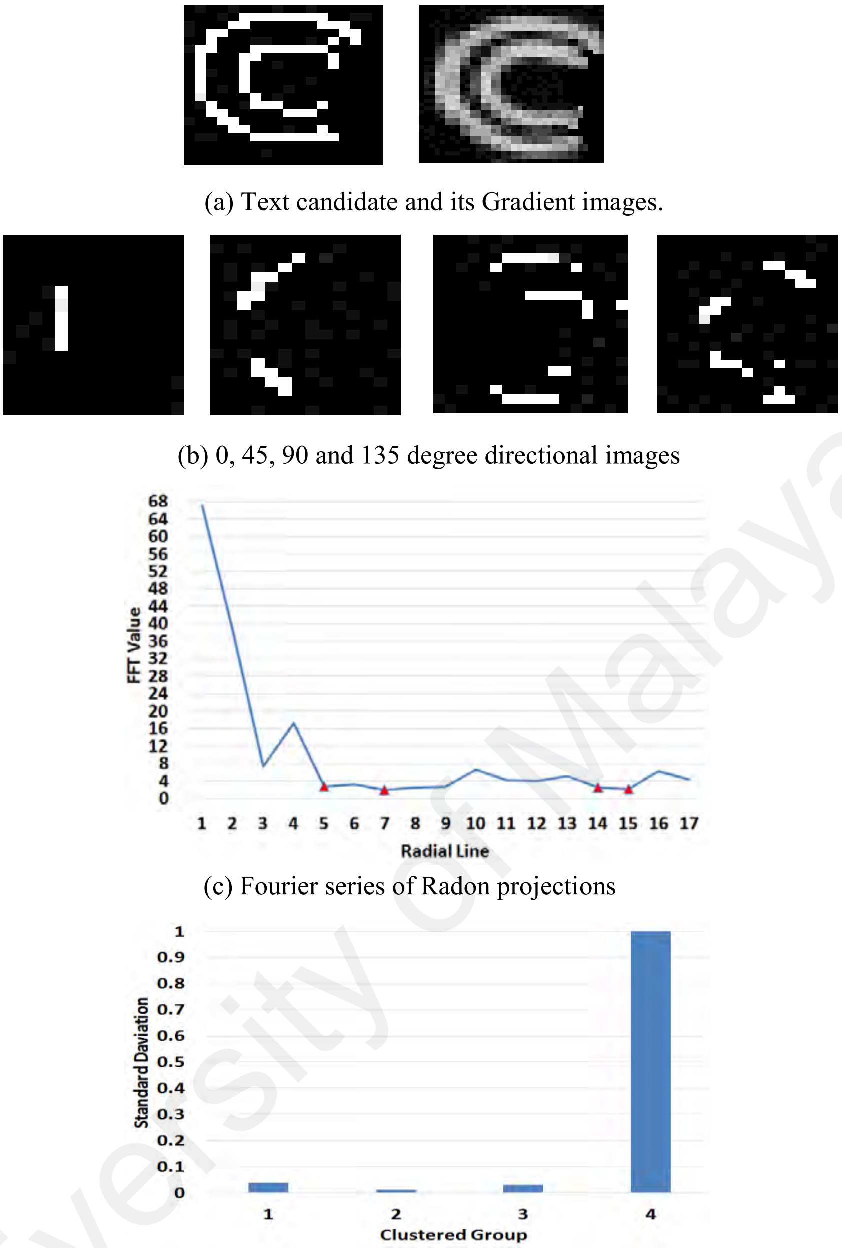
$$\text{Gradient, } |G| = \sqrt{\left(G_x^2 + G_y^2\right)} \tag{5.22}$$

$$\text{Directional Images, } G_\theta = \tan^{-1}\frac{G_y}{G_x} \tag{5.23}$$

where $\theta \in 0{,}45, 90{,}135.$

The Radon projection function $g(\rho, \emptyset)$ is the line integral of the image intensity, $(\mathrm{f(x, y)})$, that is distance ρ from the origin and at selected angle $\emptyset$ from the x-axis.

$$g(\rho, \emptyset) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)\delta(x\cos\emptyset + y\sin\emptyset - \rho)\, dy\, dx \tag{5.24}$$



(a) Text candidate and its Gradient images.

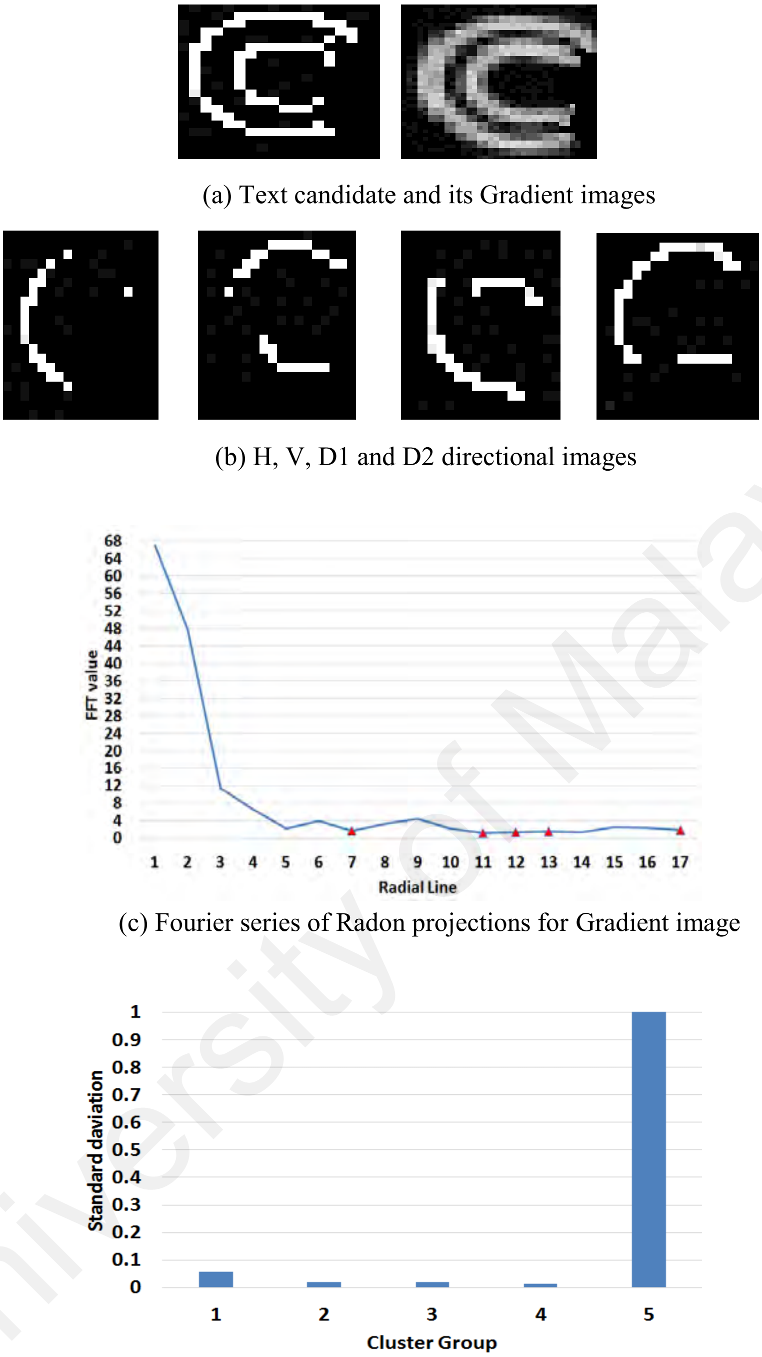
(b) 0, 45, 90 and 135 degree directional images

(c) Fourier series of Radon projections

(d) Choosing minimum standard deviation cluster

Figure 5.7: Local Features for Text Candidate Image.

To strengthen the features to extract distinct properties of text, in the same way of local features, global features is proposed for text candidates. For each text candidate image, the proposed approach uses different directional gradient masks to obtain directional images, namely, horizontal (H: 0 degree), vertical (V: 90 degree), diagonal-1(D1: 45 degree) and diagonal-2 (D2: 135 degree). In case of local features, the proposed approach checks gradient angles of the pixels in the gradient image of the text candidate image and then classifies pixels into respective directional images according to gradient angles. Therefore, such extracted features are considered as local features. In case of global features, the proposed approach convolves different directional masks with the text candidate image to obtain directional images as suggested in (Retsinas et al., 2016). The masks $[-101]$ and $[-101]^T$ are used for producing H, V, D1 and D2 directional images in this work (Retsinas et al., 2016). As a result, the same information can be seen in all the directional images unlike local features, where the union operation gives exactly the same original image. Thus, these features are considered as the global ones, which extract the global structure of text candidates. For each directional image, the proposed approach extracts feature coefficients as local features by applying Radon transform, Fourier transform and selecting the cluster which gives the minimum standard deviation. Further, the proposed approach concatenates local and global features to obtain the final feature vector for detecting text representatives.



(a) Text candidate and its Gradient images

(b) H, V, D1 and D2 directional images

(c) Fourier series of Radon projections for Gradient image

(d) Choosing minimum standard deviation cluster

Figure 5.8: Global Features for Text Candidate Image.

According to the feature extraction method as mentioned above, the features have a sufficient discriminative power for segregating non-text candidates from text. To extract such discriminative power, inspired by the work presented in (A. Zhu, Gao, & Uchida, 2016) for separating texts from the background using context information, context features is proposed based on foreground and background information of the text candidate images in this work. Since the step described in Section 5.4.1 provides foreground and background information in the form of white and black pixels, the proposed approach extracts the above feature vectors for foreground and background pixels separately. For extracting features for background, the text candidate image is complemented. It is true due to the fact that text pixels (foreground) in text candidate image have high values compared to non-text pixels (background); at the same time, the pixels of both foreground and background have almost uniform values compared to non-text candidate images. This is due to the homogeneous background for text candidate images and variations in the background for non-text candidate images. It is also true that pixel values at edges have high values compared to the others. As a result, one can expect Gaussian distribution for the features of both foreground and background for a text candidate image as shown in Figure 5.8 (a), else non-Gaussian distribution for a non-text candidate image as shown in Figure 5.9 (a). In the same way, the number of positive coefficients is larger than that of negative coefficients for foreground in case of a text candidate image as shown in Figure 5.8 (b), and vice versa in case of a non-text candidate image as displayed in Figure 5.9 (b). Therefore, context feature is defined as follows: if features of foreground and background exhibit Gaussian distributions, the image is assumed as a text candidate one, else it is a non-text one. Similarly, one more context feature is defined as if the number of positive coefficients for the foreground is more than that of background, the candidate is represented as a text, and else it is represented as a non-text candidate. The proposed approach considers the output of this step as text representatives as shown in Figure 5.10, where maximum non-text components are eliminated from the complex background image. Formally, context features are defined for detecting text and non-text using foreground and background as follows:

Context feature-1: Let say, for component C_i where, $i \in all\ components\ in\ the\ image$, the concatenate features are $f = \{f_L, f_G\}$. Then it is checked whether f follows Gaussian distribution for foreground and background. If f satisfies Gaussian distribution as defined in equation (5.25) and equation (5.26), it is represented as text candidate; else it is a non-text one.

$$f \sim \mathcal{N}(\mu, \sigma^2) \tag{5.25}$$

$$N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu^2)}{2\sigma^2}} \tag{5.26}$$

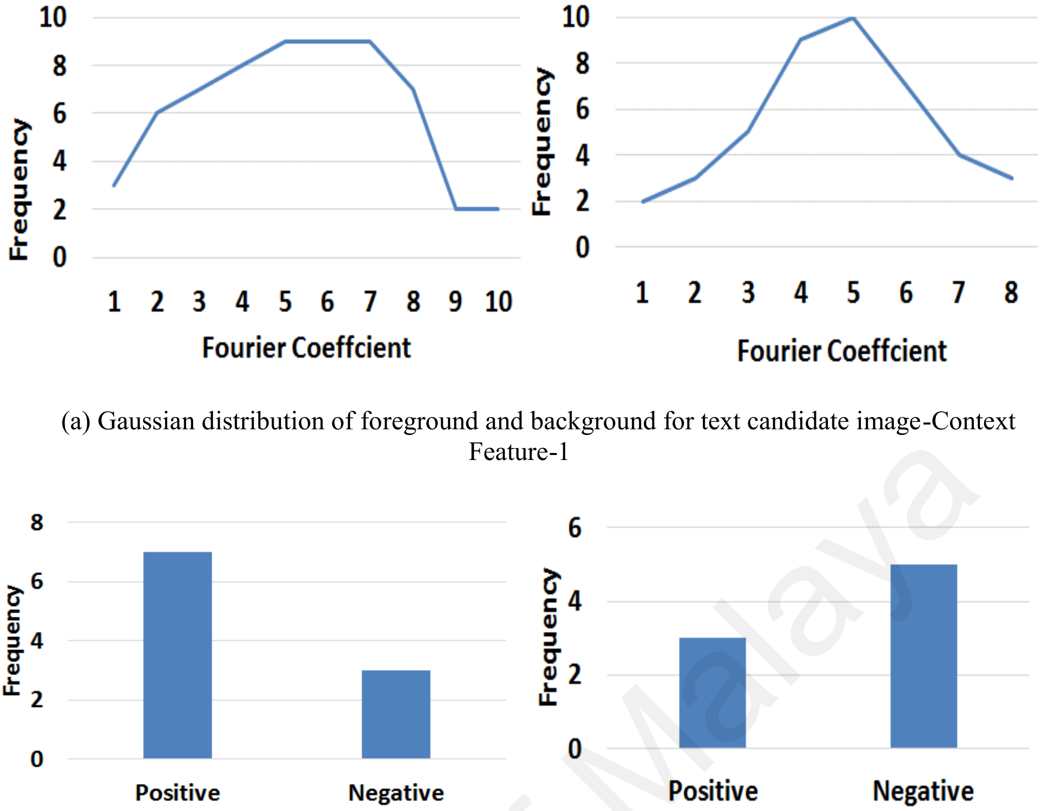
Where μ is the mean of coefficients and σ is the standard deviation of f.

Context-feature-2: The number of positive and negative coefficients is counted in feature f. If f gives more positive coefficients for foreground than background as defined in equation (5.27) and equation (5.28), it is said as a text candidate; else it is denoted as non-text one.

$$foreground_{text}: \sum_{x} (f_j > 0) > \sum_{x} (f_{x-j} < 0) \tag{5.27}$$

where, $j = 1,2, . . , x$ in coefficient in f.

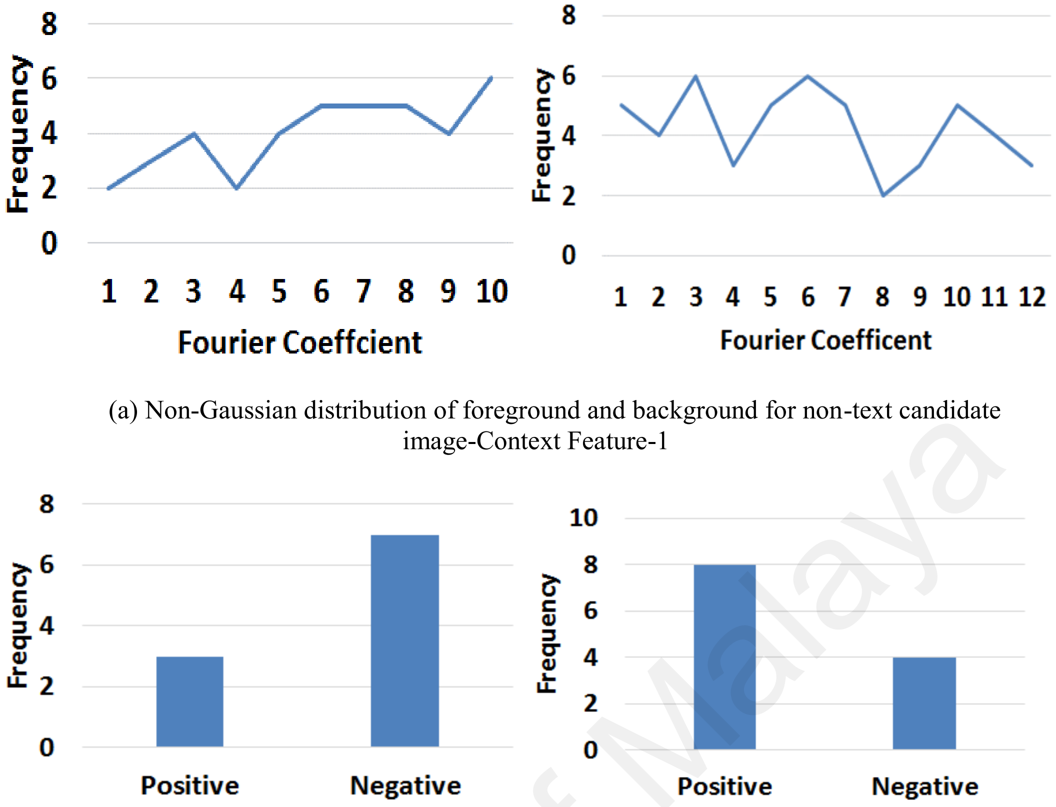
$$background_{text}: \sum_{x} (f_j > 0) < \sum_{x} (f_{n-j} < 0) \tag{5.28}$$



(a) Gaussian distribution of foreground and background for text candidate image-Context Feature-1

(b) Number of positive and negative coefficients of foreground and background for text candidate image –Context Features -2

Figure 5.9: Context feature for a text candidate image.



(a) Non-Gaussian distribution of foreground and background for non-text candidate image-Context Feature-1

(b) The number of positive and negative coefficients of foreground and background for non-text candidate image –Context Features-2

Figure 5.10: Context features for a non-text candidate image.

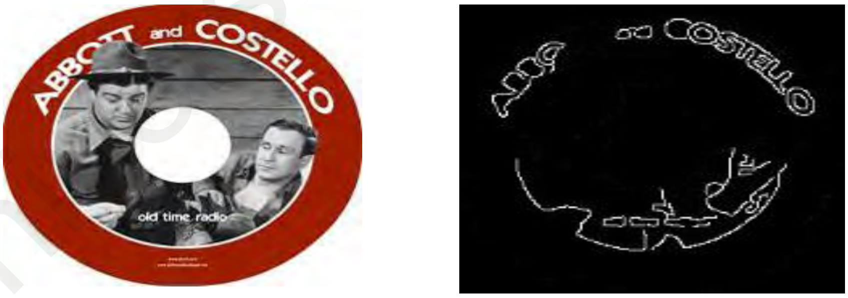


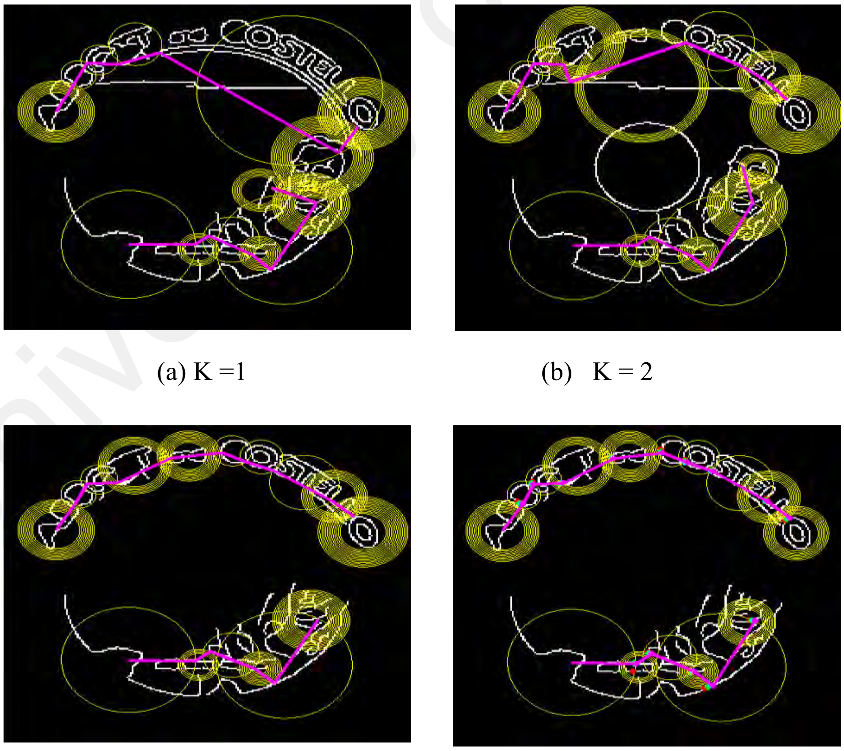
Figure 5.11: Text representatives for the input image.

5.4.3 Minimum Cost Path-based Ring Growing for Word Extraction

It is found from the text representative results shown in Figure 5.11 that these results still contain false positives and miss character components. It is believed that intensity values of text pixels have high values and are almost uniform compared to non-text pixels due to background influence (Khare, Shivakumara, Raveendran, Meng, & Woon, 2015;

Liang et al., 2015b). Besides, the spacing between character components is almost constant, and the distance between two character components is smaller than those between words or text lines (Khare, Shivakumara, Raveendran, et al., 2015; Liang et al., 2015b). Based on these observations and motivated by the work presented in (Khare, Shivakumara, Raveendran, et al., 2015) for character segmentation using least-cost path estimation and intensity values, the same minimum cost path estimation is explored to validate text pixels and ring growing for finding the nearest neighbors. At the same time, to restore missing text information, Canny edge of the input word is preferred where it can be seen that all the text information along with the background. For the components in the text representative image, the proposed approach finds minor axes and considers them as a radius to draw rings as shown in Figure 5.12, where one can see growing rings in yellow color. Then it traverses the rings to searching for the nearest neighboring white pixels of the next component. This process is continued until it gets four nearest neighbor pixels (K=4). The proposed approach compares intensity values of four neighboring pixels with the average intensity of the current component and then it chooses the neighboring pixel that gives the minimum intensity difference as a seed neighbor pixel among four neighbor pixels, which is denoted as the minimum cost path estimation. For an ideal case, the nearest neighbor should be a pixel of the adjacent character component in a text line.

However, this is not the case for the images considered in this work due to the complex background, low resolution and, curved texts. As a result, the nearest neighbor pixel can be a pixel of a non-text component. Therefore, it is expected that out of the four neighbors, at least one pixel belongs to one of the adjacent character components. Since the intensity values of non-text pixels differ from text pixels, the difference between average intensity values and non-text pixels should be higher than that between average intensity values and text pixel values. In this way, choosing K = 4 nearest neighbors helps to select correct text pixels, which are called seed neighbor pixels. The value of K is fixed by experiments, performed on the data chosen randomly from collected databases. The effect of K = 4 is shown in Fig. 10, where one can see when K =1 in Figure 5.12 (a), the line in Magenta color is not traversing along text line direction due to non-text pixels chosen by the growing process, when K =2 in Figure 5.12 (b), the line in Magenta color slightly traverses along text lines but misses few characters, when K=3 and K=4 as shown in Figure 5.12 (c) and (d), the lines in Magenta color move along text direction correctly by passing through all the correct seed neighbor pixels of character components, respectively. This shows that the value of K must be higher than 2 for achieving better results. However, according to experiments on the database, K = 4 is better than K=3 when the whole database is considered. Supporting experiments are provided in the experimental section.



(a) K =1

(b) K = 2

(c) K =3

(d) K = 4

Figure 5.12: Illustration for choosing values for K.

To restore missing text, the proposed approach grows two to three rings of the components in the text representative image, and while growing, if it finds neighbor components in the image, it continues the growing process. Otherwise, it checks at the same position in the Canny image. If any component is available, the growing process extracts the component and starts finding the next neighbor in the text representative image. Similarly, the proposed approach uses the Canny image for restoring missing information.

The above process iterates until it reaches an end of a text line and no more text representatives are found in the image. This results in the minimum cost path, which is denoted as a forward path in the text representative image by connecting seed neighbor pixels of the components in the text line as displayed in Figure 5.12 (c) and (d), where the line in Magenta color represents forward path. Based on this information, in order to remove false positives, the following properties, namely, intensity values of seed neighbor pixels, and behavior of the forward path have been defined. The proposed approach finds a backward path by applying the same process of forwarding path to find to the last component as shown in Figure 5.13 (a), where a forward path marked by Magenta color and a backward path marked by Cyan color can be seen.

For a text component, it is expected that the forward and backward paths should choose the same seed neighbor pixels with a small deviation between the paths as shown in Figure 5.13 (a), where it can be seen Magenta line and Cyan line use the same seed neighbor pixels for most of the character components. For non-text components, the forward and backward paths may use different seed neighbor pixels, which results in two different paths with a high deviation between the paths as shown in Figure 5.13 (a), where it can be noticed that Magenta and Cyan lines appear two separate lines for non-text lines. The proposed approach draws lines perpendicular to the seed neighbor pixels of the forward path to the backward path as portrayed in Figure 5.13 (b) for text lines to extract this property, and Figure 5.13 (c) for non-text lines, where it can be seen perpendicular lines between the paths. It finds the distance from the forward path to the backward path and vice versa using perpendicular lines between them. The proposed approach calculates the standard deviation for those distances values for forward paths and backward paths, respectively. If the standard deviation of forwarding path is almost the same as the standard deviation of backward path, it is called as a text component; else it is a non-text one. In Figure 5.13 (c) and Figure 5.13 (d), the deviation is low for text lines but high for non-text lines.

Formally, property-1 is defined as follows. Let the forward path $\boldsymbol{fw}$ be consist of seed neighbor pixels $t = 1,2,\ldots,f_p$ and expressed as $fw = \{fw_{p1}, fw_{p2}, \ldots.., fw_{pt}\}$. Let the backward path bw be consists of seed neighbor pixels $s = 1,2,\ldots,f_b$ and expressed as $bw = \{bw_{p1}, bw_{p2}, \ldots.., bw_{ps}\}$. For every point in fw_t, a vertical line is drawn in bw in an arbitrary position bw_a in bw line. So, the distance between two points fw_t and bw_a is calculated by Euclidian distance for the forward path as defined in equation (5.29):

$$dist_{(fw,bw)} = \sqrt{\left(fw_{pt} - bw_{pa}\right)^2} \tag{5.29}$$

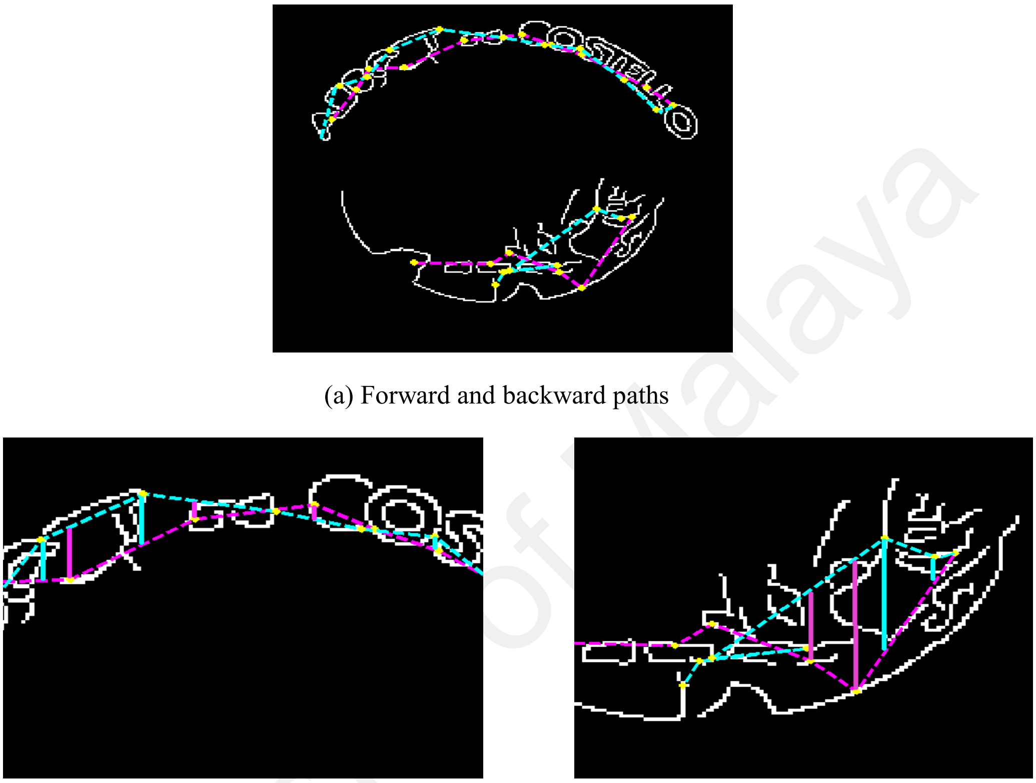
Thus, in the same way, the calculation is done for the backward path as defined in equation (5.30):

$$dist_{(bw,fw)} = \sqrt{\left(bw_{ps} - fw_{pa}\right)^2} \tag{5.29}$$

Further, separately the standard deviation of all the distances is calculated in forward fw and backward path bw. Then property-1 is stated with a certain threshold as defined in equation (5.31) and equation (5.32).

$$\left(fw_{p\_text-}bw_{p\_text} \right) \leq .5 \tag{5.31}$$

$$\left(fw_{p\_non\_text-}bw_{p\_non-text} \right) > .5 \tag{5.32}$$



(a) Forward and backward paths

(b) Property-1 for text lines

(c) Property-1 for non-text line

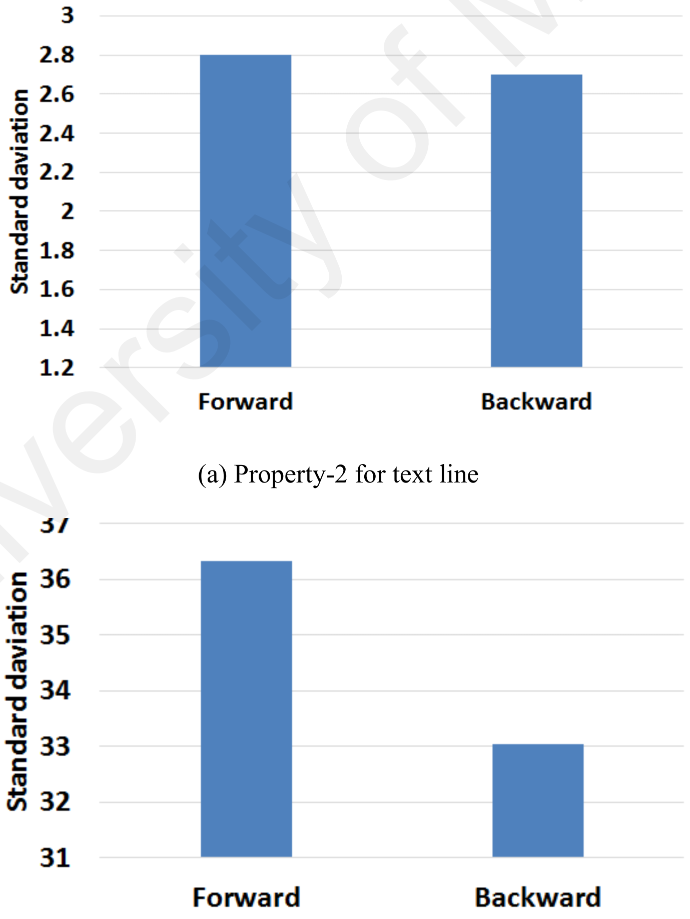
Figure 5.13: Property-1 for finding standard deviation for the distances between the paths.

In the same way of property-1, the proposed approach considers the gray difference between average intensity values of the current component and the seed neighbor pixels of forward and backward paths as discussed. Then it finds the standard deviation for those gray difference values. If the standard deviation of forward path is almost the same as the standard deviation of backward path, it is represented as a text component; else it is a non-text one. It is illustrated in Figure 5.14, where it is seen the standard deviation of gray difference values of the seed neighbor pixels of forward and backward paths are almost the same for the text line in Figure 5.13 (a), while in Figure 5.14 (b), the standard deviation of forward and backward paths differs much for the non-text line in Figure 5.13 (a). Property-2 is defined in equation (5.33) for text lines and equation (5.34) for non-text lines with a certain threshold. Here the threshold value is set as 0.5 which is determined empirically.

$$\left(fw_{g\_text} - bw_{g\_text} \right) \leq .5 \tag{5.33}$$

$$\left(fw_{g\_nontext} - bw_{g\_nontext} \right) > .5 \tag{5.34}$$

Where fw denotes forward path, bw denotes backward path, while g_text and g_nontext denote gray difference value of seed neighbor pixels of text line and non-text line, respectively.



(a) Property-2 for text line

(b) Property-2 for non-text line

Figure 5.14: Property-2 for finding standard deviation for the gray difference of forward and backward paths.

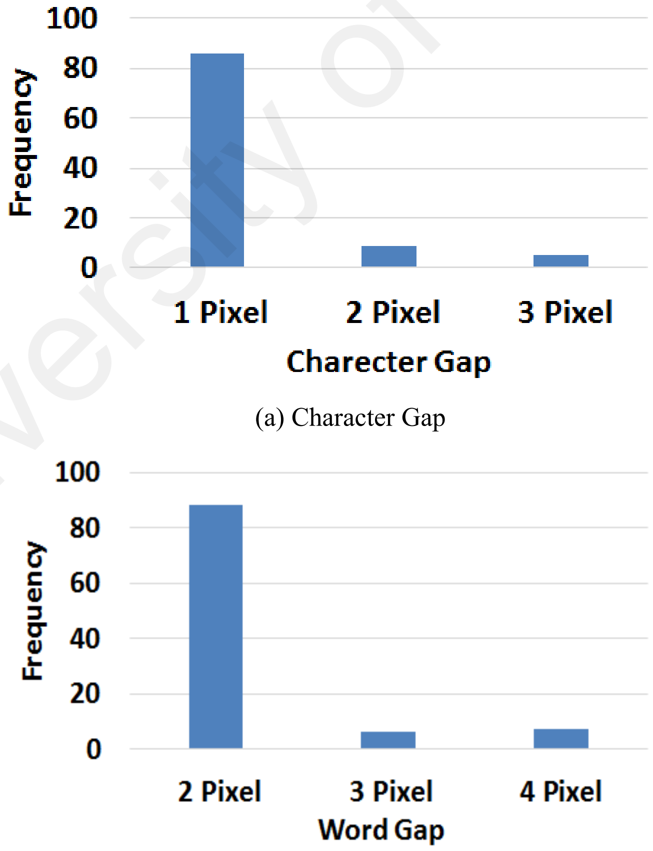
Similarly, the gray difference of the first component and its seed neighbor pixels of adjacent character components is considered as iteration-1 value. Then the gray difference of the second component with its seed neighbor pixels of adjacent character components is considered as iteration-2 value, and so on. In this way, iteration values are obtained for both forward and backward paths of text and non-text lines. After finding two iteration values, the proposed approach finds the difference between the values of iteration-1 and iteration-2. This iterative process continues until the difference of iteration values gets almost zero, which is called converging. If the iterative processes of forward and backward paths converge with a certain threshold which is found experimentally, the line is assumed as a text line; else it is assumed as a non-text one, which is defined in equation (5.35) for text line and equation (5.36) for a non-text line.

$$(convergence_{text\_fw} \& convergence_{text\_bw}) \leq 5 \tag{5.35}$$

$$(convergence_{nontext\_fw} \& convergence_{nontext\_bw}) > 5 \tag{5.36}$$

If a resorted line satisfies the above three properties, the proposed approach considers the candidate text as a true text positive; else it is a false positive.

In order to determine the thresholds for character gap, word gap, and end of lines, 100 text lines are chosen from databases randomly for experimentation. As noted from the ring growing process presented in Figure 5.12 that when the growing process finds seed neighbor pixels of an adjacent character with the current character component, the distance from the current character component and the adjacent character can be considered as character gap, and in the same way it is noted that word gap is always larger than character gap. Therefore, the proposed approach applies this ring growing process on 100 text lines to determine a valid character and word gaps. The threshold is considered for the end of lines as larger than the average of word gaps in text lines automatically. In case the text line does not contain more than one word especially for license plate images, the proposed approach considers the threshold as larger than the average of character gaps. To verify the distance between character and word components, histogram operation is performed on the distances between character components and words components given by ring growing process for 100 text lines as shown in Figure 5.15, where it is noted that one pixel distance contributes for the highest peak as a character gap as shown in Figure 5.15 (a), and two pixels distance contributes to the highest peak as a word gap represented in Figure 5.15 (b). Figure 5.15 depicts that the distance between characters is often one pixel, while the distance between words is two pixels. Sample word detection results for different images can be viewed in Figure 5.16, where the proposed approach detects words well irrespective of orientation, fonts, font size, and background.



(a) Character Gap

(b) Word Gap

Figure 5.15: Illustrating distance between characters and words in a text line.

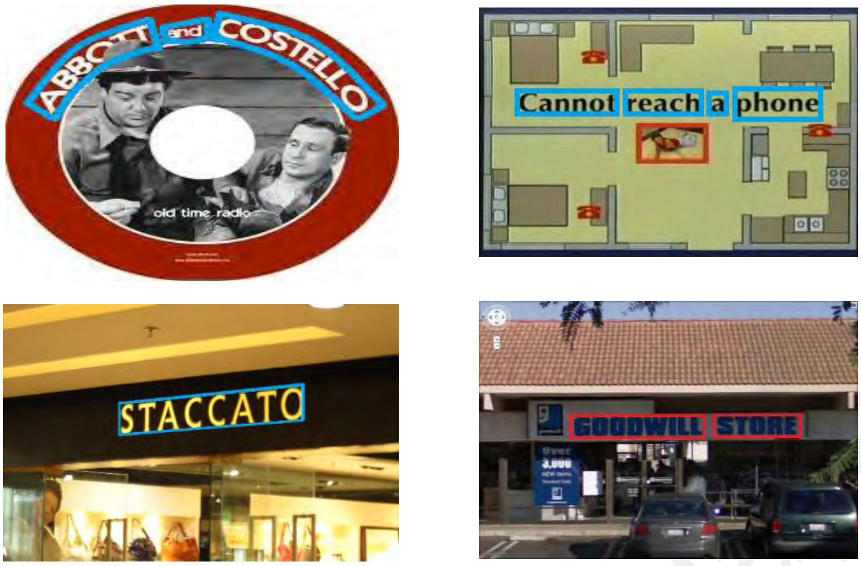


Figure 5.16: Sample word detection results of the proposed approach for different databases.

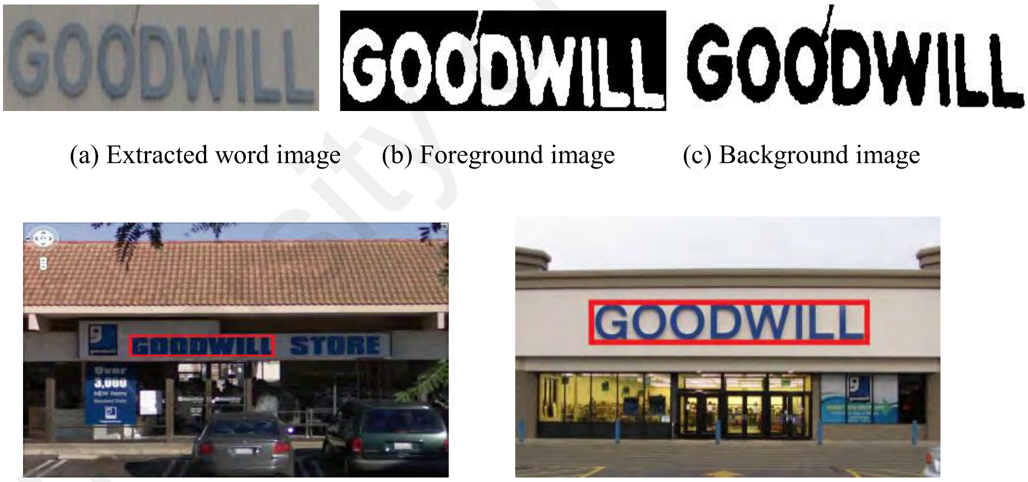
5.4.4 Foreground and Background based Features for Word Spotting

For each word given by the previous step, the proposed approach extracts local and global features as described in Section 5.4.2 for foreground and background information separately. Since the step presented in Section 5.4.1 separates foreground and background information, the same step has been used for detected words to extract local and global features, which helps in identifying the relation between foreground and background information of character components for word spotting as shown in Figure 5.17. The detected word image is shown in Figure 5.17 (a). Figure 5.17 (b) and Figure 5.17 (c) depict the separation of foreground and background information. It is noted from Figure 5.17 (c) that background information is obtained by complementing the foreground image in Figure 5.17 (b). Inspired by the work presented in (Retsinas et al., 2016) for keyword spotting in handwritten document images using distance measures defined in equation (5.37), the same distance measures is preferred to use for word spotting in this work. This is because it is shown in (Retsinas et al., 2016) that the proposed distance measure can cope with the challenges caused by different handwriting styles and variations as the proposed problem poses font or font size variations, distortions, multiple orientations, etc.

The distance $d(test, word\ )$ is measured between query image *test* and a word image by following the equation:

$$d(test\ ,word\ ) = \frac{\|test1 - word\ 1\|_2}{N1} + \frac{\|test2 - word\ 2\|_2}{N2} \tag{5.37}$$

Where $test1$, $word1$ are feature matrices generated by the global feature, while $test2$, $word2$ are the feature matrices generated by local features. N1 and N2 are the lengths of global and local feature matrices, respectively. Sample results for spotting words in different images are shown in Figure 5.17 (d) for the query word “GOODWILL”.



(a) Extracted word image (b) Foreground image (c) Background image

(d) Word is spotted successfully

Figure 5.17: Sample word spotting of the proposed approach.

In summary, an integrated approach based on fractional means and context feature has been proposed for word spotting. In beginning fractional means has been explored for detecting text candidates in input images, which also provides foreground and background information of images. Then Radon and Fourier transforms are investigated for the text candidates detected by fractional means features in gradient domain to extract features locally and globally. Based on the features, the proposed method extracts context information with the help of foreground and background. Furthermore, a new minimum cost path estimation has been proposed based on ring growing for restoring missing information during the detection of text representatives by moving along text direction to extract words. For the detected words, the proposed method extracts the above mentioned local and global features for both foreground and background to perform keyword spotting.

5.5 Experimental Results

This section gives detail description of the dataset, evaluation metrics, and comparative methods. In section 5.5.1, dataset and evaluation metrics of word detection and spotting are discussed. Section 5.5.2 gives results of word detection method. Here detection accuracy is compared between before candidate region detection and after candidate region detection. The result of keyword spotting method is presented in Section 5.5.3.

5.5.1 Dataset and Evaluation

In this sub-section, dataset and evaluation metrics used for detection and spotting are discussed.

5.5.1.1 Word/Bib number Detection in Images

For multi-modal based text detection approach, a dataset of size 200 images of sports, Olympics, Marathon and running race has been created, which is named as “collected datasetI” (CDI). The standard dataset of size 217 images has also been used, which is called the RBNR dataset (Ben-Ami et al., 2012). In total, the proposed technique has been tested on 417 images. Well-known measures such as R, P, and F have been used for evaluating the performances. For recognition, recognition rate (RR) has been used for evaluating recognition results. The definition and instructions as proposed in (Ben-Ami et al., 2012) are followed for both text detection and recognition experimentation in this work. Since the RBNR dataset provides the ground truth for calculating measures, the same ground truth has been used for the experiments on RBNR data and for CDI, measure has been calculated manually because there is no ground truth.

To evaluate the proposed technique, text detection, and recognition experiments are conducted before text candidate region (TCR) detection and after TCR detection. For text detection experiments, Shivakumara et al. (Palaiahnaktoe Shivakumara et al., 2010) technique which is suitable as discussed in Section 5.2 and Epshtein et al. (Epshtein et al., 2010) which is well-known text detection approach in scene images have been chosen. In the same way, Roy et al. (S. Roy et al., 2012) which proposes wavelet and gradient combination for binarizing text lines of video, Moghaddam et al. (Moghaddam & Cheriet, 2010) which proposes a multi-scale adaptive binarization technique for degraded images, Wolf et al. (Wolf et al., 2002) which propose a method for binarizing multimedia documents, Chattopadhyay et al. (Chattopadhyay et al., 2013b) which propose automatic selection of binarization techniques for improving recognition results, and Howe (Howe, 2013) which proposes an automatic way to tune parameters and threshold values to improve binarization results, have been chosen for calculating recognition rate using Tesseract OCR engine (Smith, 2007). The proposed technique in this work consists of face + skin + text detection by (Palaiahnaktoe Shivakumara et al., 2010)+ binarization by (Howe, 2013) + Tesseract OCR (Smith, 2007) .

5.5.1.2 Keyword Spotting in Video

Similarly, for keyword spotting, a database has been created, named as collected databaseII (CDII) from the TRECVID video database as there is no standard database. Words have been collected from 12 different video streams and also from ICDAR 2013 video and natural scene (Karatzas et al., 2013), which includes noisy words, oriented words, different contrasts, fonts and font sizes. This gives a total of 1200 words database.

The recall, precision, and f-measure are defined as in equation (5.38), equation (5.39) and equation (5.40), respectively, to evaluate the performance of the method. The total number of the instances (N) is counted manually for every word as ground truth. Recall (R) is defined as the total number of correctly matched word (CW) divided by the total words (NW) of every word. Precision (P) is defined as the total number of correctly matched keywords divided by the total number of matched words (MW). More details can be found in (Khare, Shivakumara, Raveendran, et al., 2015; Liang et al., 2015b). The evaluation metrics are defined as follows.

$$Recall\ (R) = \frac{CW}{NW} \tag{5.38}$$

$$Precision\ (P) = \frac{CW}{MW} \tag{5.39}$$

$$fmeasure(F) = \frac{2\times(R\ \times P)}{(R+P)} \tag{5.40}$$

For Context based approach, in addition to ICDAR 2013 (which is used for texture-spatial based keyword spotting approach), other standard databases, namely, ICDAR 2015 (Karatzas et al., 2015), YVT (Nguyen et al., 2014) and NUS (Liang et al., 2015b) video data which generally suffer from low resolution, small fonts or arbitrary orientations, ICDAR 2015 (Karatzas et al., 2015), ICDAR 2013 (Karatzas et al., 2013), SVT (K. Wang, Babenko, & Belongie, 2011b) and MSRA natural scene data (Yao et al., 2012) which suffer from complex background, font variations, font size variations or arbitrary orientations, and UCSD (Zamberletti et al., 2015), Medialab (Zamberletti et al., 2015) and Uninsubria (Zamberletti et al., 2015) License plate data which suffer from non-uniform illumination, vehicle movements, headlight, etc. have been considered. In case of video dataset, since the proposed approach requires frames for word spotting, keyframes have been extracted from videos of different databases as follows: 599 frames from 24 video of ICDAR 2015 (I15), 1300 frames from 15 videos of ICDAR 2013 (I13), 200 frames from YVT (YT), 100 images from NUS (NS), which include horizontal, non-horizontal and curved text line images. That is, total 2199 frames are used for experimentation. In case of natural scene datasets, randomly 400 images from ICDAR 2015 (I15), 67 images from ICDAR 2013 (I13), 65 images from SVT (ST) and 100 images from MSRA (MA) have been considered, which gives total 632 images for experimentation. In case of License plate datasets, 890 images from UCSD (UD), 70 images from Medialab (MB) and 376 images from Uninsubria (UA), that is, total 1336 images are considered for experimentation. In total, 4167 images and frames are considered to ensure that the dataset covers wide variations.

The proposed approach consists of two main steps, namely, word detection and word spotting. Therefore, well-known measures (Khare, Shivakumara, Raveendran, et al., 2015; Liang et al., 2015b), namely, R, P, and F are considered or examining word detection step. The measures suggested in (Retsinas et al., 2016), namely, Precision@5 (P@5), Mean Average Precision (MAP), Normalize Discounted Cumulative Gain (NDCG), and Binary Normalized Discounted Cumulative Gain (BNDCG) are considered for evaluating the word spotting step. The definitions of the above-mentioned measures for word detection are as follows.

Precision@5 (P@5) is defined as a ratio of the number of retrieved relevant images to the total number of retrieved irrelevant and relevant images:

$$\textit{Precision@5} = \frac{\text{The number of retrieved relevant images}}{\text{The total number of retrieved irrelevant and relevant images}} \quad (5.41)$$

Mean Average Precision (MAP) is defined as a mean of the average precision scores for each query:

$$MAP@5 = \frac{\sum_{q=1}^{m}(Pr(query) \times rel(q))}{\text{The number of retrieved relevant images}} \tag{5.42}$$

where $rel(q)$ is 0 and 1. The value '1' corresponds to relevancy of word with query word, and the value '0' denotes non-relevancy at rank q.

The Normalize Discounted Cumulative Gain (NDCG) introduces a penalty if most pertinent words appear at the bottom in the retrieval list.

$$NDCG = \frac{DCG}{IDCG} \tag{5.43}$$

$$DCG = rel_1 + \sum_{h=2}^{m} \frac{rel_h}{\log_2 h} \tag{5.44}$$

where $\boldsymbol{rel_h}$ is the relevance judgment at position $\boldsymbol{h}$**,** whereas, IDGG is the ideal DCG denotes the ground truth.

Binary Normalize Discounted Cumulative Gain (BNDCG) is equal to NDCG except that binary relevance, i.e., either 1 or 0. More details for the word spotting measures can be found in (Retsinas et al., 2016) because the same evaluation scheme has been followed in this work.

For spatial-texture based keyword spotting approach, existing method (S. Lu & Tan, 2007), which is developed for keyword spotting in camera document images using the characteristics of connected components and shape codes, is tested. Since there are no methods reported for keyword spotting in the video in literature, Lu and Tan's method (S. Lu & Tan, 2007) has been implemented to show that document-based spotting method might not work well for videos.

To show the superiority of the context-based approach, the states of art approaches of document spotting work have been implemented for comparative studies. Howe (Howe,

2013) proposed Inkball models for handwritten document images, which combines different character localization approaches for word spotting. Retsinas et al. (Retsinas et al., 2016) explored projections of oriented gradients, which explores Radon and Fourier transform for feature extraction to perform word spotting. Mondal et al. (Mondal et al., 2016) proposed a flexible sequence matching technique with a learning free approach for word spotting in degraded handwritten document images, which proposed improved sequential matching to improve the results. In addition to these approaches, the proposed spatial-texture based method has also been evaluated. Shivakumara et al. (Shivakumara, Liang, Roy, Pal, & Lu, 2015) proposed an approach for keyword spotting video images, which explores texture and spatial based features. The reason to use the approaches (Mondal et al., 2016; Retsinas et al., 2016; Shivakumara et al., 2015) proposed for keyword spotting in handwritten document images is that these are the state of art methods and robust to degradations, distortion, handwriting styles, contrast variations, etc. as the proposed problem has variations in fonts, font size, orientations, etc. In addition, these approaches show that conventional approaches, such as the approaches proposed for printed document images, may not work well for handwritten documents. However, all the approaches do not focus on different video images, natural scene images and license plate images as these are much more complex than handwritten images and video images considered (K. Wang & Belongie, 2010).

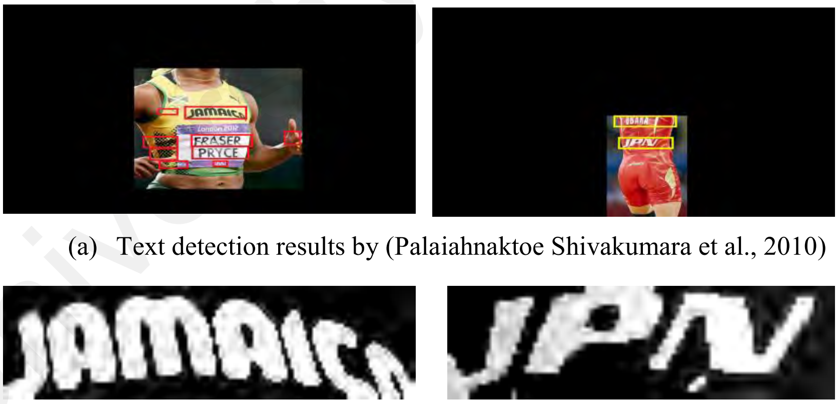
For calculating measures of word spotting in this work, at least five repeated words are considered as query words in respective databases, and then the proposed approach runs the same query word at least five times to perform keyword spotting. Further, the average of ten times retrieved results are reported in the following tables.

5.5.2 Experiments for Word Detection

In this sub-section experimental result of multi-modal based approach has been discussed. The influence of biometric feature on text candidate detection is tested. The text detection result before applying biometric feature and after applying biometric feature has been compared. The performance of the proposed method is compared to against existing methods.

5.5.2.1 Experiments for Multimodal based Approach

Sample results by the biometric feature for the text candidate regions in Figure 5.1 are shown in Figure 5.18 (a), where it can be noticed that the text detection technique detects texts well. Similarly, sample binarization results for the text lines detected by the text detection technique are displayed in Figure 5.18 (b). Finally, the binarized results feed into the OCR engine to recognize texts or bib numbers as depicted in Figure 5.18 (b). The OCR results are displayed within in quotes.



(a) Text detection results by (Palaiahnaktoe Shivakumara et al., 2010)

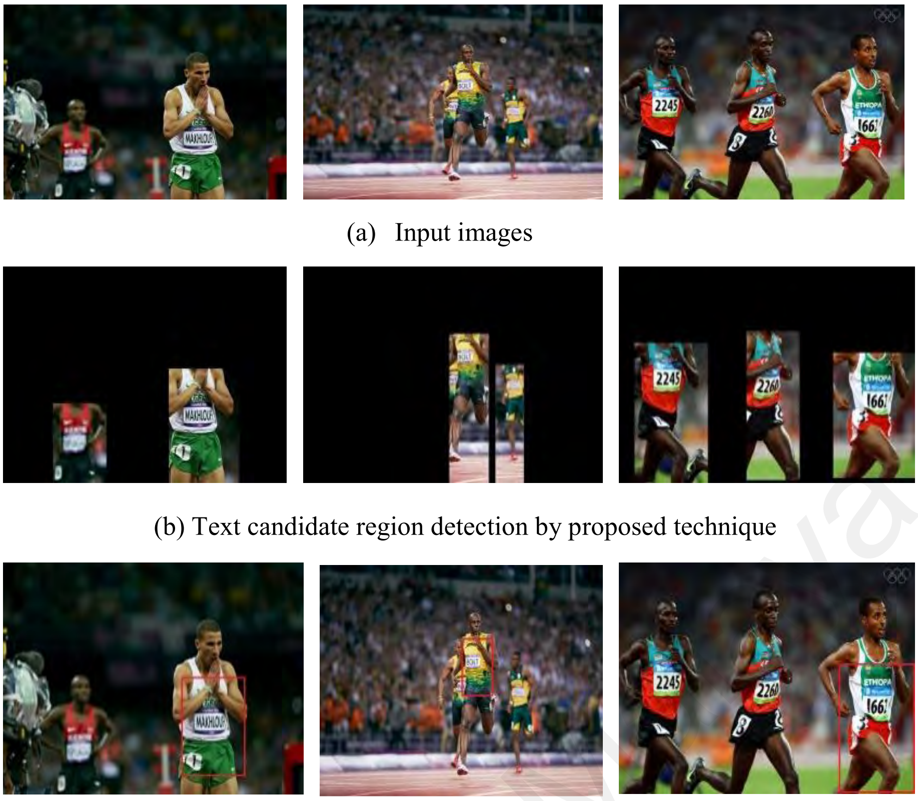
“JI\lN\\g” “JPN”

(b) Recognition results by (Howe, 2013)

Figure 5.18: Text detection and recognition results of text candidate region.

5.5.2.2 Comparative Study and Discussion

To show the effectiveness of the TCR detection step in multi-modal technique, TCR detection results are compared with the results of the RBNR technique in terms of R, P, and F. Sample images of the proposed and the existing methods are given in Figure 5.19. From this figure, it is observed that the proposed step detects TCR well for the input images in Figure 5.19 (a) as shown in Figure 5.19 (b) compared to the results in Figure 5.19 (c) given by the existing technique. Figure 5.19 (c) shows that the existing technique misses some of the TCR due to low contrast and blur, while the proposed technique detects well. The quantitative results depict that the performance of the proposed technique is better than the existing technique in terms of R and F-measure for both the two datasets. However, the precision for the RBNR technique is higher on CDI than that of the proposed technique due to low false positives. The proposed technique sometimes produces false positives when face + skin combination fails to detect proper TCRs. However, overall, the proposed technique outperforms the existing technique. The main advantage is that proposed method uses both face and skin while the existing technique uses only face information. As a result, the existing method misses several text candidates. Therefore, the recall for the existing method is lower than the proposed method.



(a) Input images

(b) Text candidate region detection by proposed technique

(c) Text candidate region detection by RBNR technique (Ben-Ami, Basha, & Avidan, 2012)

Figure 5.19: Sample text candidate region detection of proposed and existing techniques on collected data.

Table 5.1: Performance of the proposed and existing techniques for text candidate region detection.

Methods	RBNR Data			Our Data		
	R	P	F	R	P	F
Proposed method	78.26	58.06	66.66	70.85	84.73	77.17
(Ben-Ami et al., 2012)	49.15	97.38	65.32	57.44	76.33	65.55

Table 5.2: Performance of the text detection techniques before and after text candidate region detection.

Text Detection Methods	Before Text Candidate Region Detection						After Text Candidate Region Detection					
	RBNR Data			Our Data			RBNR Data			Our Data		
	R	P	F	R	P	F	R	P	F	R	P	F
Shivakumara et al. (Palaiahnaktoe Shivakumara et al., 2010)	83.31	17.39	24.95	38.61	37.55	34.89	24.3	40.23	24.52	44.23	46.57	41.09
Epshtein et al. (Epshtein et al., 2010)	20.1	1.88	3.09	5.277	1.95	2.61	45.02	31.25	33.33	13.98	7.51	8.91

Table 5.2 shows that text detection techniques give better results at R, P, and F-measure after TCR detection compared to before TCR detection. However, it can be observed from Table 5.2 that the method described in (Palaiahnaktoe Shivakumara et al., 2010) gives a good recall for RBNR before TCR detection, but gives a poor recall for the same RBNR data after TCR detection. This is because sometimes TCR detection may lose text lines during text candidate region detection. The precision of Shivakumara et al.'s method is better for RBNR data after TCR. For CDI, both the techniques generate better results after TCR detection compared to before TCR detection in terms of R, P, and F-measure. Sample qualitative results of the proposed technique on CDI are shown in Figure 5.20, where it can be seen that the proposed technique detects almost all the bib numbers and texts in the TCR. Therefore, it can be concluded that text detection techniques alone are not good enough to solve the problems of identifying bib/text in running race images. Hence, TCR detection is useful to improve the performances of existing text detection techniques.

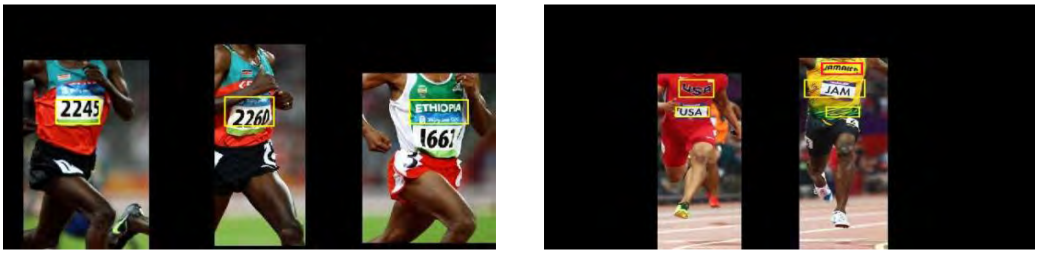


Figure 5.20: Sample text detection results of the proposed method on collected data.

Experiments are also conducted on recognition through binarization techniques to show the usefulness of TCR detection using the multimodal way. Table 5.3 and Table 5.4 show the recognition rates of different binarization techniques before TCR and after TCR on CDI and RBNR data, respectively. To show that the conventional OCR engine is not suitable for text recognition in the marathon and running race images before TCR detection, the whole input image is sent to OCR engine to calculate recognition rate. Then text lines detected by the text detection technique (Palaiahnaktoe Shivakumara et al., 2010) from the whole input image are passed to OCR engine to calculate recognition rate to show the improvements compared to the whole image. Since Shivakumara et al.'s method gives better results for text detection after TCR detection as shown in Table 5.2, the same technique has been chosen for bib number or text detection experiments for recognition. Similarly, the whole TCR is sent to OCR directly to calculate recognition rate after TCR detection. Again, text lines detected by the text detection technique after TCR are sent to OCR. Table 5.3 and Table 5.4 show that for both the datasets, different binarization techniques give poor results for the whole image before TCR detection compared to those after TCR detection. It is noted from Table 5.3 that Howe technique gives a better recognition rate after TCR compared to other binarization techniques for CDI. Table 5.4 shows that Roy et al.'s method gives a better recognition rate for the RBNR data after TCR compared to the other techniques. However, overall, when results after TCR on both RBNR and CDI are compared, it is found that the recognition rate for CDI is lower than that of RBNR data. This shows that CDI is much more complex data compared to the existing RBNR data.

Further, it is found from Table 5.5 that the performance of proposed technique yields better results than the state-of-the works in terms of R and F-measure for both the datasets. However, the precision of the existing technique on CDI is higher than the proposed technique. This is because the way the existing technique converts RGB to binary and uses

Tesseract for recognition is different from proposed technique. The main reason to get better results is the same as discussed in text detection experiments. This is valid because the dataset contains several images where faces are not visible, which is common in case of marathon and running race images.

Table 5.3: Character recognition rate of the binarization techniques on CDI dataset (in %).

Binarization Methods	Before-TCR		After-TCR	
	The whole image	Text	TCR	Text
(S. Roy et al., 2012)	2.1	34.2	4.3	42.2
(Moghaddam & Cheriet, 2010)	2.4	25.3	4.1	27.2
(Wolf et al., 2002)	2.7	12.2	3.3	16.3
(Chattopadhyay et al., 2013b)	1.8	16.2	3.1	21.3
(Howe, 2013)	3.1	31.6	5.3	43.3

Table 5.4: Character recognition rate of the binarization techniques on RBNR dataset (in %).

Binarization Methods	Before-TCR		After-TCR	
	The whole image	Text	TCR	Text
(S. Roy et al., 2012)	2.3	20.0	3.9	49.0
(Moghaddam & Cheriet, 2010)	3.1	17.2	2.1	25.3
(Wolf et al., 2002)	2.6	10.2	2.7	19.0
(Chattopadhyay et al., 2013b)	1.3	12.2	2.4	20.2
(Howe, 2013)	3.3	18.2	4.1	33.8

Table 5.5: Character recognition rate of the proposed and existing techniques (in %).

Methods	RBNR Data			Our Data		
	R	P	F	R	P	F
Proposed method	55.3	76.5	64.19	49.69	56.16	52.73
(Ben-Ami et al., 2012)	52.0	68.0	58.6	28.48	77.04	41.58

5.5.3 Experiments on Keyword Spotting in Video

In this sub-section, the experiments performed for evaluating keyword spotting approaches have been discussed.

5.5.3.1 Experiments for Texture-Spatial Features based Approach

The proposed word segmentation method in texture-spatial based approach has been evaluated by conducting experiments on the diversified words dataset, which includes

1047 horizontal video and 125 text lines natural scene images (Karatzas et al., 2013) from ICDAR 2013, 325 non-horizontal video text lines, 70 arbitrary oriented video text lines. In total, 1567 text line images are considered for testing the word segmentation method. The quantitative results of the word segmentation method are tabulated in Table 5.6. It is noticed from the table that the method is the best for horizontal text lines, the lowest for curved data because segmenting words from curved like circles or semi-circles is not as easy as horizontal text lines.

Table 5.6: Performance of the word segmentation method on different datasets.

Dataset	R	P	F
ICDAR 2013	0.85	0.90	0.87
Curved dataset	0.72	0.78	0.75
Non-horizontal Straight lines	0.82	0.85	0.83
Horizontal Straight Line	0.88	0.92	0.90

5.5.3.2 Experiments for Fractional Means based Approach

The proposed Fractional Means based approach introduces few key steps to achieve better results for word spotting in video/natural/license plate images. To understand the effectiveness of each step, 599 frames of ICDAR 2015 video are considered for experimentation. Since the primary focus of the proposed work is to spot the keywords in ICDAR 2015 standard video data available publicly, ICDAR 2015 video frames are considered as representative frames for all the other databases considered in this work. In Section 5.4.2 the proposed approach extracts local and global features for detecting representatives. Word spotting has been performed using only local features, only global features and using both. The results reported in Table 5.7 shows that the results using only local features and using only global features give low results compared to using both local and global features. Therefore, it can be concluded that both local and global features play a role in achieving the good results. In Section 5.4.3, the proposed approach uses K = 4 neighbors for word detection. When the values for K =1, K = 2, K =3, K = 4 are analyzed, the proposed approach scores poor results for K = 1, K =2 and good results for K = 3 and K = 4 according to Table 5.7. This shows that K = 3 and K = 4 help to identify correct seed neighbor pixel of the character component despite noisy background pixels present near to text line. In the same Section 5.4.3, three new properties have been introduced to remove false positives. Word spotting results are reported in Table 5.7 for various experiments to show the effectiveness of each property. According to Table 5.7, each property contributes to false positive elimination. However, compared to the results of all three properties together, results of individual properties are low. In Section 5.4.4, the proposed approach extracts features using the foreground and background for word spotting. Experimental results on foreground features alone and background features alone show that the results of both foreground and background scores better results than an individual. Therefore, it can be concluded that all the key steps contribute effectively in achieving better word spotting results for the different types of images.

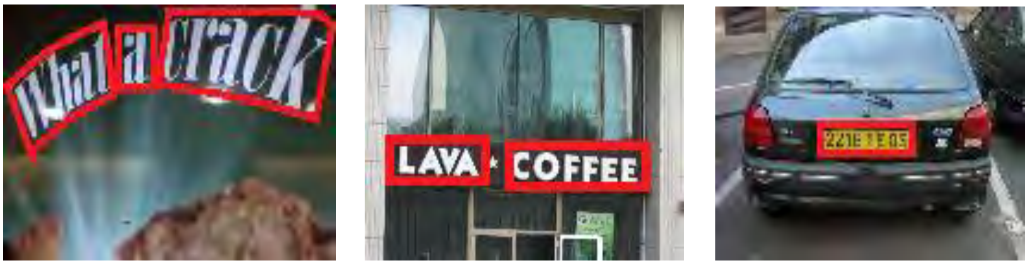
Table 5.7: Performance of the key steps of the proposed approach.

Key Steps of the Propose Approach		P@5	MAP	BNDCG	NDCG
Proposed Approach with Local Features Only		0.53	0.32	0.62	0.63
Proposed Approach with Global Features Only		0.62	0.45	0.62	0.63
Proposed Approach with both Local and Global Features		0.65	0.47	0.71	0.72
Proposed Approach - K = 4 Nearest Neighbors	K=1	0.47	0.34	0.42	0.42
	K=2	0.46	0.38	0.54	0.54
	K=3	0.63	0.48	0.72	0.72
	K=4	0.65	0.47	0.71	0.72
Proposed Approach with three properties of false positive removal	With Property-1.	0.55	0.47	0.61	0.62
	With Property-2.	0.53	0.46	0.52	0.52
	With Property-3.	0.47	0.31	0.54	0.54
	With all three	0.65	0.47	0.71	0.72

Table 5.7: Continued

Key Steps of the Propose Approach	P@5	MAP	BNDCG	NDCG
Proposed Approach with Foreground Features only	0.61	0.36	0.64	0.65
Proposed Approach with Background Features only	0.32	0.23	0.41	0.41
Proposed Approach with both Foreground and Background Features	0.65	0.47	0.71	0.72

For word detection, qualitative results of the proposed approach are shown in Figure 5.21. The proposed approach detects words well in all the three type images despite multi-orientation, complex background, and degradations, especially in license plate images as shown in Figure 5.21. Quantitative results of the proposed approach are reported in Table 5.8, which shows that the proposed approach scores almost same for ICDAR 2015, ICDAR 2013 especially F-measure and slightly poor results for NUS. The reason is that NUS data contains images of circle shaped text lines, which may cause a problem to segment words. Similarly, for natural scene databases, the proposed approach scores F-measure almost all the same results. However, for MSRA, the recall is lower than precision due to many Chinese text lines images with arbitrary-orientation according to Table 5.8. In the same way, for the license plate databases, the F-measure is low for the UCSD data than the other two datasets because UCSD dataset contains many images affected by severe blur and contain too small font text. Overall, when the performance of the proposed approach is compared among three databases, the result for license plate images is lower than the other two databases. This is due to severely blurred images and too small font of license plate images.



Video Scene License Plate

Figure 5.21: Sample word detection results of the proposed approach for different databases.

Table 5.8: Performance of the proposed approach on word detection.

Database	Video				Natural Scene				License Plate		
Measures	I15	I13	YT	NS	I15	I13	ST	MA	UD	MB	UA
Recall	0.70	0.68	0.66	0.62	0.68	0.70	0.65	0.56	0.39	0.58	0.61
Precision	0.76	0.76	0.73	0.69	0.74	0.77	0.74	0.63	0.44	0.64	0.65
F-Measure	0.72	0.71	0.69	0.65	0.70	0.73	0.69	0.59	0.41	0.60	0.62

Sample results for keyword spotting in different videos are shown in Figure 5.22, where it can be noted that for the query words affected by different causes such as background variation, poor resolution, and blurring, the proposed approach performs keyword spotting well. Therefore, the proposed approach can withstand the poses created by different types of video.

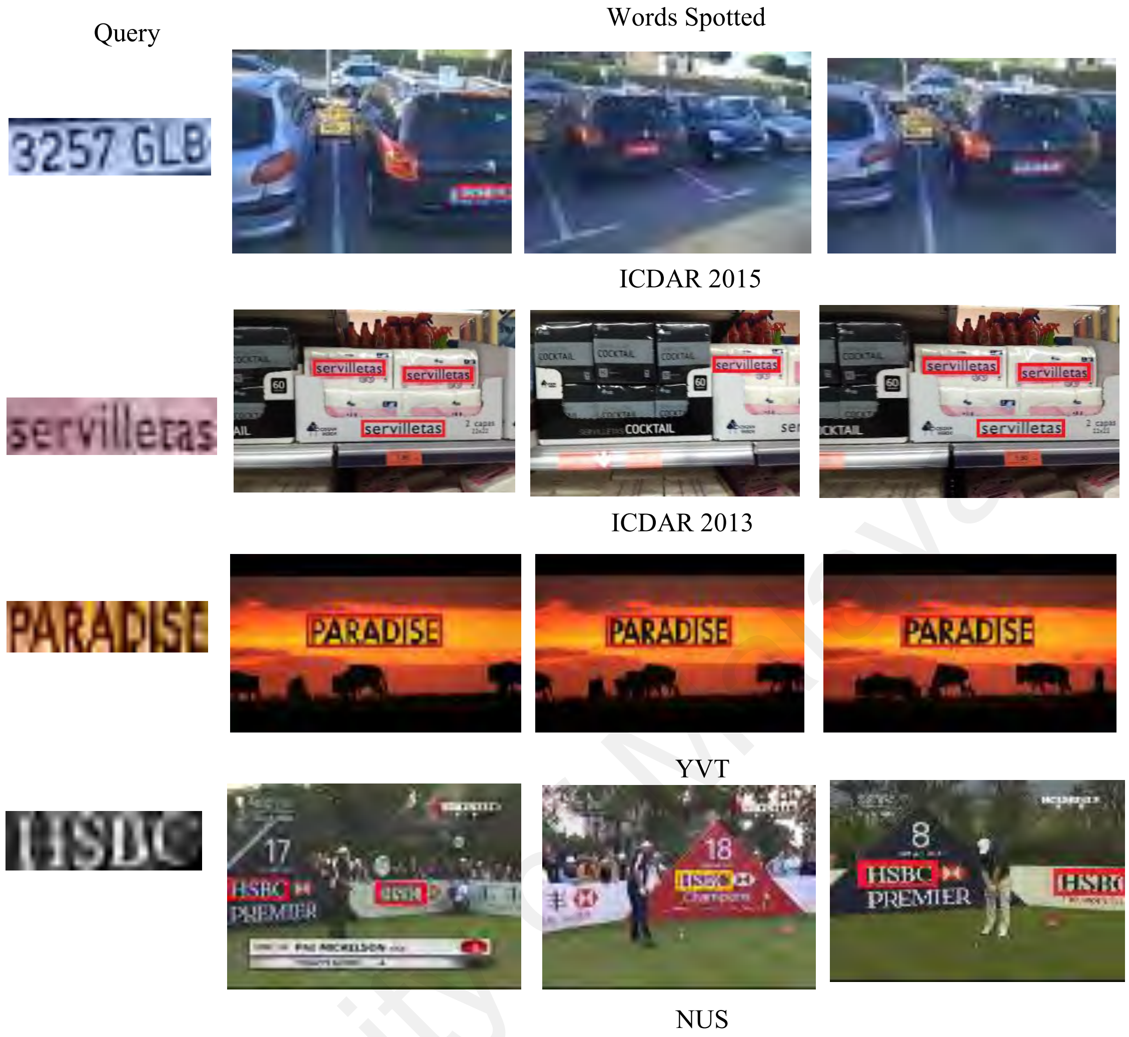


Figure 5.22: Sample keyword spotting results of the proposed approach on different videos.

A few sample results in different natural scene databases are shown in Figure 5.23, where it is noted that for the different query words, the proposed approach spotted exact words from images with the complex background, oriented texts, blurred text, etc.

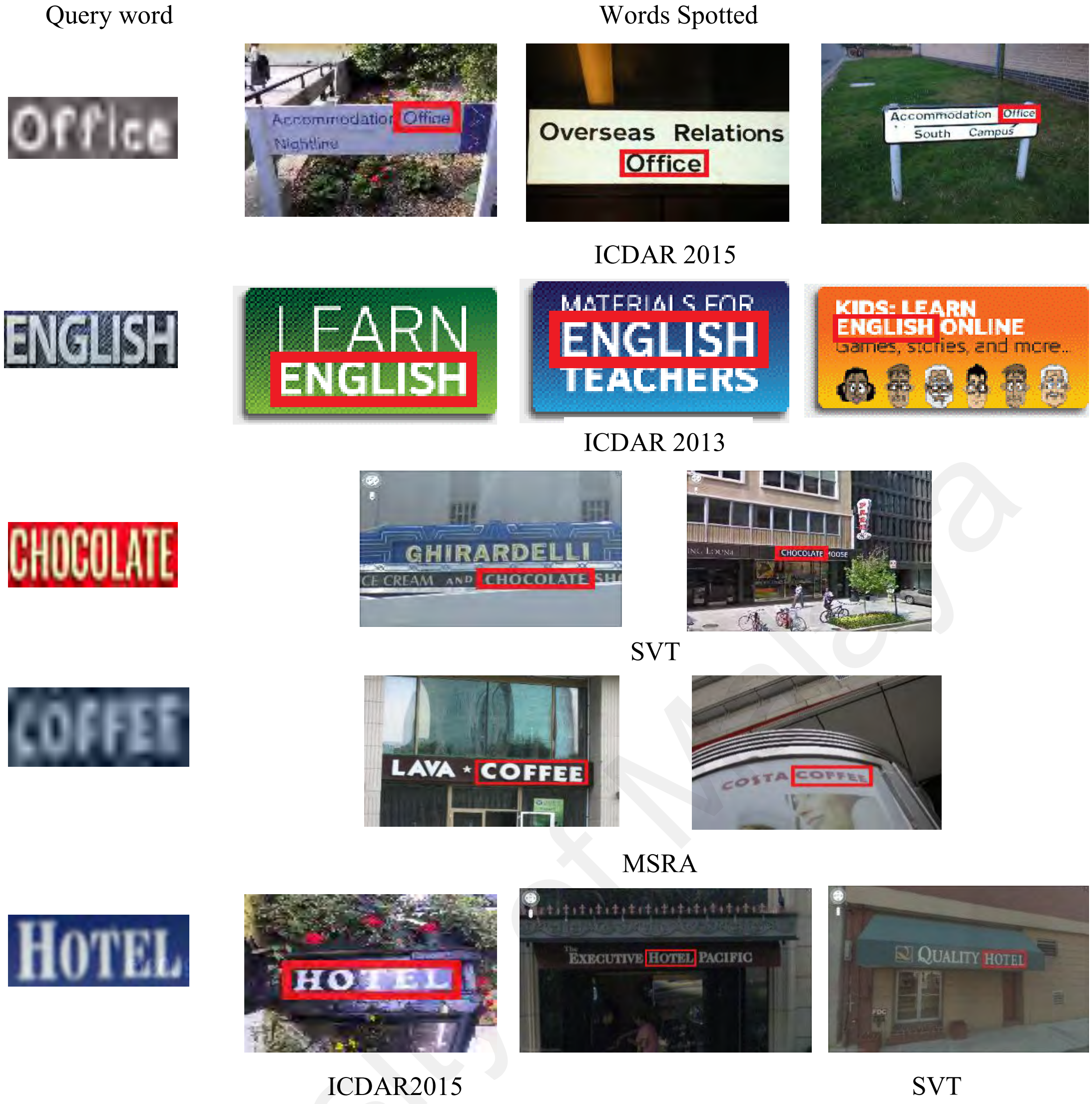


Figure 5.23: Sample keyword spotting results for different natural scene datasets.

Sample results of the proposed approach for keyword spotting in different license plate images are shown in Figure 5.24. In Figure 5.24, one can see that for the query words with blurred and, illumination affection, the proposed approach performs spotting well on different databases. This shows that the proposed approach can withstand the challenges posed by different license plate databases.

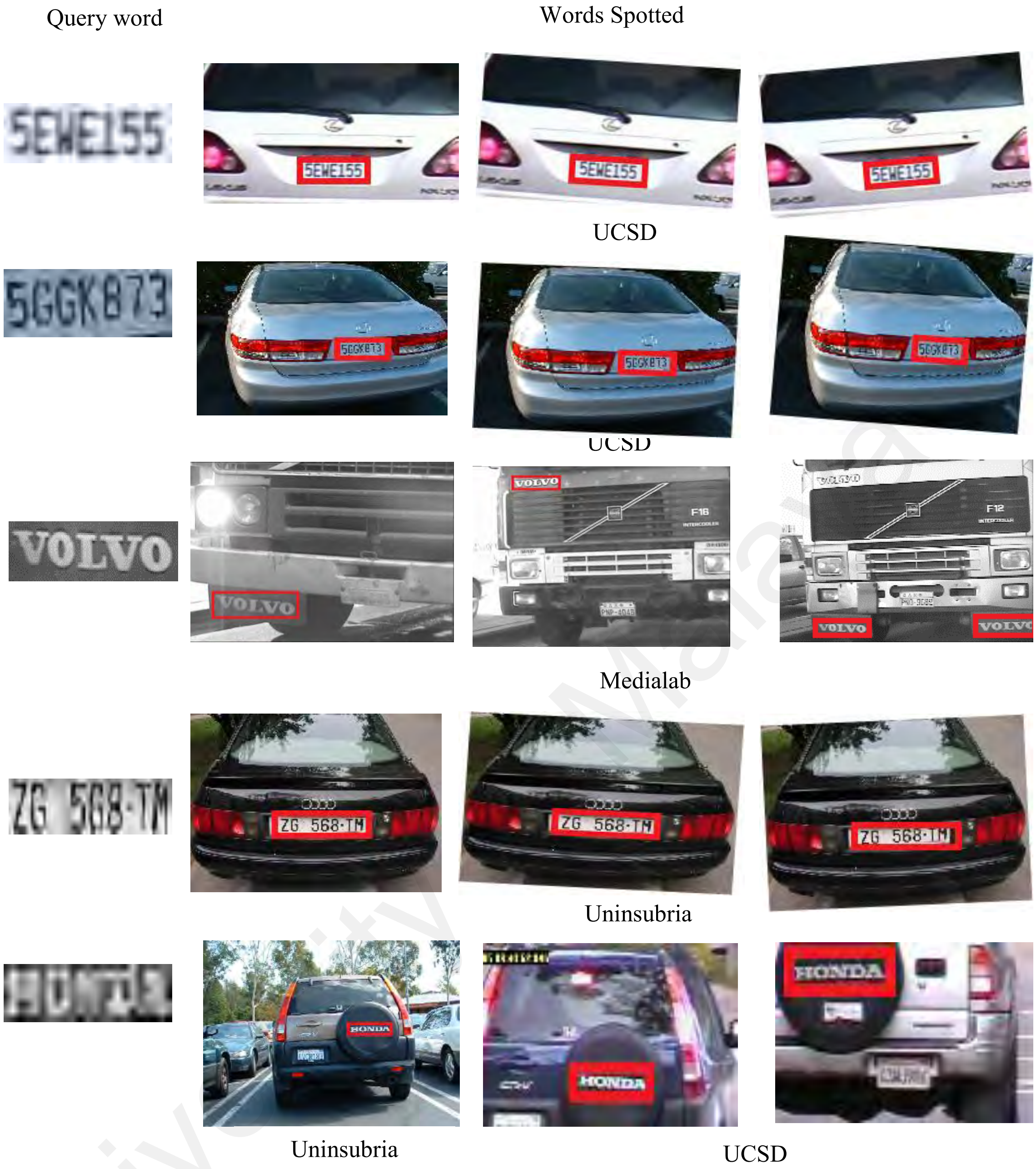
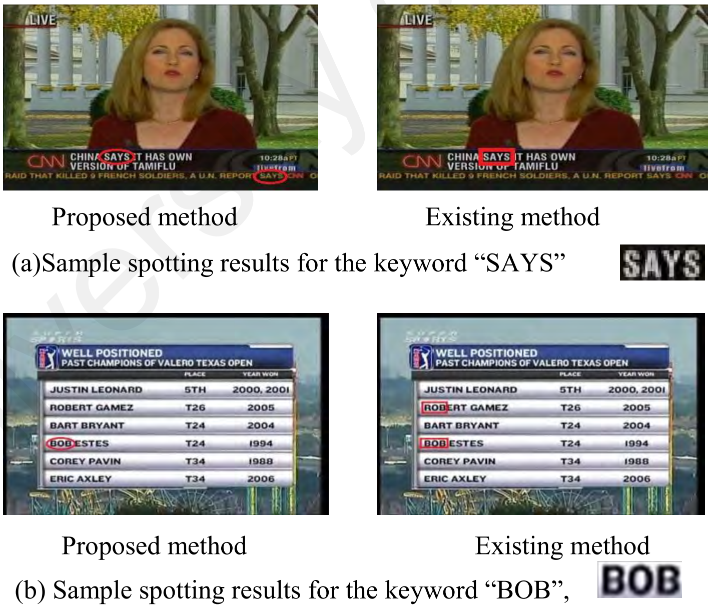


Figure 5.24: Sample keyword spotting result of the proposed approach on different License Plate databases.

5.5.3.3 Comparative Study and Discussion

Sample keyword spotting results of the texture-spatial based approach and existing method are shown in Figure 5.25 for the query words “SAYS and “BOB”, respectively, in Figure 5.25 (a) - Figure 5.25 (b). The spotted words are marked by an oval with red color in video frames. For the word “SAYS”, the proposed method has spotted two words in the frame correctly as shown in Figure 5.25 (a), where one can observe that out of these two words, one word has good contrast, and the other has low contrast. On the other hand, the existing method has spotted one word which has good contrast but fails to spot the low contrast word as shown in Figure 5.25 (a). This shows that existing method does not work well for low contrast images. Similarly, one more example is shown in Figure 5.25 (b), where for the query word “BOB”, the proposed method has spotted the word “BOB” correctly in the video frame as shown in Figure 5.25 (b), while the existing method has spotted a correct word “BOB” and one more falsely spotted word “ROB” as both the word shapes look almost similar. Since the existing method uses shape code for keyword spotting, sometimes it may fail to distinguish the words that have almost similar shapes. Therefore, the current existing method gives poor R, P, and F-measure for CDII experimentation. However, the proposed method performs better compared than state of the art because of the advantage of global and local features.



Proposed method Existing method

(a)Sample spotting results for the keyword “SAYS”

Proposed method Existing method

(b) Sample spotting results for the keyword “BOB”,

Figure 5.25: Qualitative results of the proposed and existing methods.

The cutoff distance *d* which is required for cosine distance metric is set as 0.0010 for matching the query word with the words in the database for keyword spotting. The threshold value is determined based on experimental results. For graphs of recall, precision and f-measure vs. different cut of distances have been plotted for the query word "WEATHER" as shown in Figure 5.26, where it is noted that the recall is increasing as the cutoff distance increases, while the precision and f-measure are decreasing gradually. Therefore, the cutoff distance is set as 0.0010 as a tradeoff between recall and precision. The same value has been used for experimentation in this work.

Table 5.9 tabulates the quantitative results of the proposed and existing work. This shows that the proposed method for keyword spotting generates better results in terms of R, P, and F-measure than the existing method (S. Lu & Tan, 2007). The main reason for obtaining low accuracy is that the existing method was developed for high contrast camera-based images but not for video.

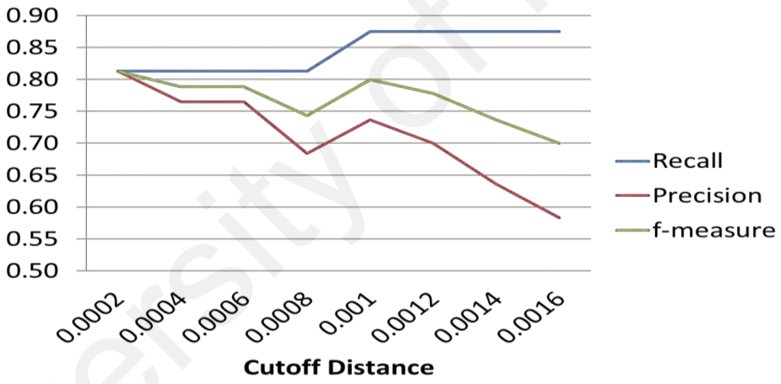


Figure 5.26: Tradeoff for cutoff distance on "WEATHER" keyword.

Table 5.9: Performance of the proposed and existing methods on keyword spotting.

DATA		Proposed –TSF			(S. Lu & Tan, 2007)		
Query Word	N	R	P	F	R	P	F
AIRPORT	13	1	0.87	0.93	0.62	0.57	0.59
CHRYSLER	11	1	0.69	0.81	0.64	0.58	0.61
LAYOFFS	11	0.91	0.71	0.8	0.55	0.43	0.48
NASCAR	13	0.69	0.69	0.69	0.38	0.36	0.37
PANDA	9	0.89	0.8	0.84	0.78	0.54	0.64

Table 5.9: Continued

DATA		Proposed –TSF			(S. Lu & Tan, 2007)		
	33	0.67	0.63	0.65	0.58	0.50	0.54
	8	1	1	1	0.88	0.58	0.70
	11	1	0.85	0.92	0.73	0.62	0.67
	15	1	0.56	0.71	0.80	0.71	0.75
	16	0.88	0.74	0.8	0.69	0.58	0.63
Total	140	0.90	0.75	0.81	0.64	0.54	0.59

Quantitative results of the context-based approach and existing approaches are reported in Table 5.10, which shows that the proposed approach is the best at all measures except P@5 compared to the existing approaches. Since the existing approaches are capable of handling degradations and handwriting variations, the approaches extract relevant words which fall in the top five words, while the proposed approach aims at spotting exact relevant words at first and thus sometimes it loses accuracy. However, it is noted from Table 5.10 that the precisions of both the proposed and existing approaches are lower than recall and the other measures. This is due to matching, which helps in finding relevant words but sometimes fails to find exact words due to the complex background.

Quantitative results of the proposed and existing approaches are listed in Table 5.10 for different videos. Experiments have been conducted on individual databases and all together as reported in Table 5.10 . Experimental results show that the difference between individual databases and all together is not marginal. It shows the proposed approach is consistent with different situations. It is also observed from Table 5.10 that the performances of existing approaches are poor compared to the proposed method in terms of all the measures. This is because the existing approaches are developed for specific databases and applications, while the proposed approach is developed for different types of databases and applications. Table 5.10 depicts that the NDCG, BNDCG, and recall is better than precision for the proposed method. Therefore, the proposed approach is good in spotting relevant words but misses to spot exact words sometimes, and hence its precision is lower than other measures.

Quantitative results of the proposed and existing approaches are reported in Table 5.10, which shows that the proposed approach outperforms the existing approaches in terms of recall, precision, and relevancy. However, it is noted from Table 5.10 that precision is lower than the other measures in case of the proposed approach. This is due to matching and complexity of the images. Therefore, overall, the performance of the proposed approach in spotting keywords on license plate images is not high compared to that of word spotting in document analysis. The main reason is shown in Figure 5.25, where it can be seen that the query words are affected by severe blur, non-uniform illumination, and font variations, the spotted word may not be the exact query word. Therefore, there is a scope for future work on these issues to improve the results.

Table 5.10: Performance of the proposed and existing approaches for keyword spotting on different videos, natural scene and license plate databases.

Measures	Methods	Video					Natural Scene					License Plate			
		I15	I13	YT	NS	All	I15	I13	ST	MA	All	UD	MB	UA	All
P@5	(Howe, 2013)	0.53	0.55	0.43	0.33	0.46	0.61	0.58	0.51	0.55	0.56	0.48	0.46	0.51	0.48
	(Retsinas et al., 2016)	0.61	0.56	0.55	0.42	0.53	0.59	0.64	0.54	0.51	0.57	0.49	0.52	0.55	0.52
	(Mondal et al., 2016)	0.60	0.56	0.53	0.38	0.51	0.57	0.61	0.41	0.44	0.50	0.40	0.49	0.53	0.47
	(Shivakumara et al., 2015)	0.47	0.51	0.40	0.27	0.41	0.53	0.51	0.38	0.42	0.46	0.31	0.42	0.39	0.37
	Proposed	0.68	0.64	0.72	0.57	0.59	0.63	0.74	0.52	0.54	0.60	0.53	0.61	0.62	0.58
MAP	(Howe, 2013)	0.31	0.32	0.29	0.14	0.26	0.37	0.35	0.34	0.33	0.34	0.31	0.32	0.39	0.34
	(Retsinas et al., 2016)	0.47	0.46	0.42	0.31	0.41	0.46	0.53	0.41	0.42	0.45	0.34	0.39	0.42	0.38
	(Mondal et al., 2016)	0.42	0.31	0.36	0.21	0.32	0.35	0.36	0.21	0.24	0.29	0.31	0.38	0.45	0.38
	(Shivakumara et al., 2015)	0.22	0.28	0.23	0.15	0.22	0.35	0.31	0.20	0.21	0.26	0.19	0.29	0.27	0.25

Table 5.10: Continued

Measures	Methods	Video					Natural Scene					License Plate			
		I15	I13	YT	NS	All	I15	I13	ST	MA	All	UD	MB	UA	All
	Proposed	0.71	0.62	0.69	0. 47	0.62	0.63	0.74	0.52	0.54	0.60	0.42	0.46	0.47	0.45
NDCG	(Howe, 2013)	0.61	0.54	0.41	0.20	0.44	0.42	0.53	0.48	0.46	0.47	0.51	0.46	0.59	0.56
	(Retsinas et al., 2016)	0.62	0.66	0.63	0.39	-	0.56	0.59	0.59	0.51	0.56	0.52	0.59	0.61	0.57
	(Mondal et al., 2016)	0.60	0.62	0.61	0.17	0.50	0.50	0.59	0.44	0.33	0.46	0.41	0.60	0.60	0.53
	(Shivakumara et al., 2015)	0.52	0.58	0.42	0.18	0.42	0.36	0.51	0.39	0.39	0.41	0.29	0.35	0.51	0.38
	Proposed	0.78	0.78	0.69	0.61	0.71	0.63	0.67	0.71	0.62	0.65	0.55	0.66	0.64	0.61
BNDCG	(Howe, 2013)	0.60	0.54	0.42	0.23	0.44	0.43	0.56	0.49	0.46	0.48	0.51	0.42	0.60	0.51
	(Retsinas et al., 2016)	0.67	0.63	0.61	0.49	0.6	0.55	0.62	0.48	0.51	0.54	0.49	0.51	0.63	0.54
	(Mondal et al., 2016)	0.63	0.62	0.59	0.21	0.51	0.52	0.61	0.45	0.38	0.49	0.42	0.54	0.61	0.52
	(Shivakumara et al., 2015)	0.58	0.51	0.41	0.22	0.43	0.47	0.51	0.41	0.32	0.42	0.39	0.38	0.53	0.43
	Proposed	0.75	0.71	0.72	0.58	0.69	0.62	0.69	0.74	0.66	0.67	0.54	0.60	0.64	0.59

5.6 Summary

Overall, in this chapter, novel text retrieval methods, i.e., text detection and spotting from a different type of data, like the natural scene, video and license plate images have been investigated. In text detection, face and skin features are explored in a new way of identifying text candidate regions for input natural images. Then for each text candidate region, detection and recognition methods are used. To retrieve desired information without recognizing text, word spotting is explored using two approaches. The first approach is based on a combination of spatial texture feature whereas second is based on fractional means along with a combination of Radon and Fourier coefficients. The approaches' main advantage lies in the ability to work with variant style and size of font and some extent to distortion.

CHAPTER 6: CAPTION AND SCENE TEXT TYPES CLASSIFICATION IN VIDEO

6.1 Background

In the previous chapter, the methods for spotting keyword in video and images presented. This work spots the keywords irrespective of text type in the video where it can be seen caption and scene type text. To improve recognition performance, this chapter proposes a method for classification of caption and scene text. Generally, caption text is edited/superimposed, which results in artificially created text compared to other frame contents. This fact has been exploited in proposed approaches to identify the type of text in video frames by exploring the advantages of Discrete Cosine Transform (DCT) coefficients and wavelet coefficient.

6.2 Tampered Feature-based Approach for Line-wise Caption and Scene Text Classification

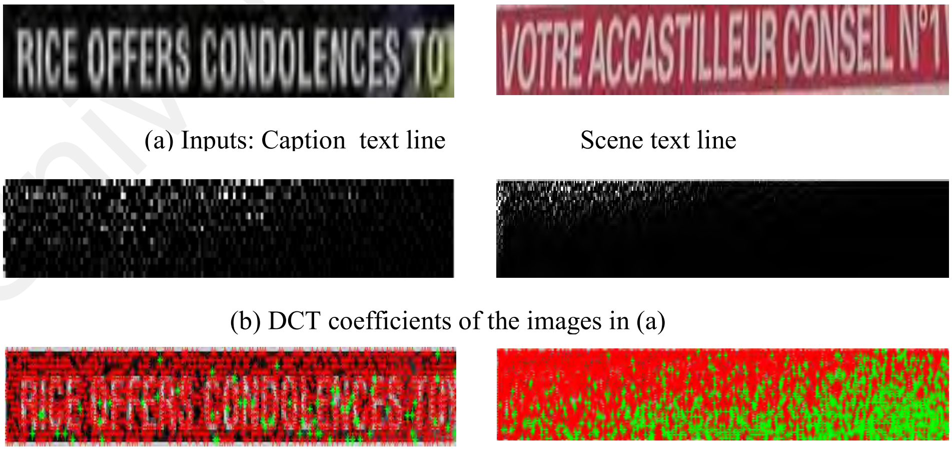
Based on literature survey on text detection in video (Khare, Shivakumara, & Raveendran, 2015; Liang et al., 2015a; Q. Ye & Doermann, 2015), one recent method is found in (Liang et al., 2015a), which gives reasonable results in spite of the inconsistency for video and scene texts detection. In addition, this method does not have any limitation on orientation. Therefore, this method is used for text line detection rather than cropping text lines manually. If the method does not detect full-text lines or misses text lines, manual segmentation is done from frames. However, it is found that text detection methods detect texts regardless of caption and scene text types with inconsistency, but they don’t have the ability to identify them (Liang et al., 2015a; P Shivakumara, N Vinay Kumar, et al., 2014; J. Xu et al., 2014). Therefore, it is necessary to develop a method for differentiating them to improve recognition rates because these two text types differ in quality, clarity, and contrast.

For this approach, text lines have been extracted following the method described in (Liang et al., 2015a). Then, this extracted text line is used as input for classification. Thus the scope of the present work is to identify texts as either caption or scene. The identification of texts can be used for video text frame classification as video frames having only caption texts, only scene texts and both caption-scene texts. This is because in general, a single frame in the video may contain caption, scene and both caption and scene texts (Liang et al., 2015a; Q. Ye & Doermann, 2015). As a result, the classification of frames is important to improve text detection results for video as discussed in Chapter 3.

It is true that caption text in the video frame is superimposed text and hence it can be considered as tampered text (Hooda, Kathuria, & Pankajakshan, 2014), while scene text can be natural text as it is a part of an image. This cue motivated to explore DCT coefficients (Haodong Li, Luo, & Huang, 2015; Zhong, Zhang, & Jain, 2000). Therefore, distribution of zero and non-zero coefficients (positive values) is explored over text line images to identify them. To extract such cues, lines are obtained according to the distribution of zero and non-zero coefficients over text line image. A new idea of studying the linearity and smoothness property of the lines has been proposed based on checking the centroid of a line falling on the line itself or not iteratively. If the centroid falls on the same line, it is considered as straightness property or else cursiveness property. The relationship between the line that represents zero coefficients and the line that represents non-zero coefficients is defined as rule-1 for classifying text types. Since it is a complex classification problem and one idea may not be sufficient, one more idea has been proposed that counts crossing points where the principal axis crosses over the actual coefficient line to study the smoothness of lines with respect to zero and non-zero coefficients (positive values). The proposed method finds the relationship between these two lines to define rule-2 for classifying them. Further, both rule-1 and rule-2 are combined to achieve better results.

6.2.1 Tampered Feature-based Approach for Line-wise Caption and Scene Text Classification

For the text lines shown in Figure 6.1 (a) as caption and scene texts, the proposed method obtains DCT images as shown in Figure 6.1 (b) for the corresponding images in Figure 6.1 (a). It is noted from the DCT images in Figure 6.1 (b) that for caption texts, high DCT coefficients are scattered over the image, while for scene texts, DCT coefficients are clustered at the top left corner of the image. The same observation can be confirmed using the distribution of zero-coefficients over images from Figure 6.1 (c), where zero coefficients are denoted by green color, and non-zero coefficients are denoted by red color. It is seen from Figure 6.1 (c) that for caption texts, zero coefficients are scattered over the image while for scene texts, dense zero coefficients gradually increase towards the bottom right corner. This cue leads to the extraction of tampered information for classifying scene and caption text in frames.



(a) Inputs: Caption text line Scene text line

(b) DCT coefficients of the images in (a)

(c) Distribution of zero and non-zero coefficients for Caption and Scene text line images in (b). Green represents zero and red represents non-zero coefficients.

Figure 6.1: DCT coefficients distribution for Caption and Scene text line images.

To extract such cues given by DCT coefficients, a window operation has been performed in a non-overlapping way over text lines as shown in Figure 6.2 (a) for caption texts, where the height of texts is considered as the width to define window size. Since this work considers text lines as an input, it also provides also the direction of text lines, therefore moving the window in arbitrary orientations is not an issue. For each window, the percentages of zero and non-zero coefficients (positive values) are computed as defined in equation (6.1) and equation (6.2), respectively. The percentage calculation makes the method invariant to different dimensions of text lines. The effect of the distribution of percentage values for the whole text line can be seen in Figure 6.2 (b), where it is noted that for caption texts, the distributions of zero coefficients (red color bars) and non-zero coefficients (blue color bars) do not have uniform variations, while for scene texts, both the coefficients have uniform variations. There is a gradual change in the percentage of zero coefficient (red color bars) values as the window moves over text lines in case of scene texts and almost the same variations for non-zero coefficients (blue bars).

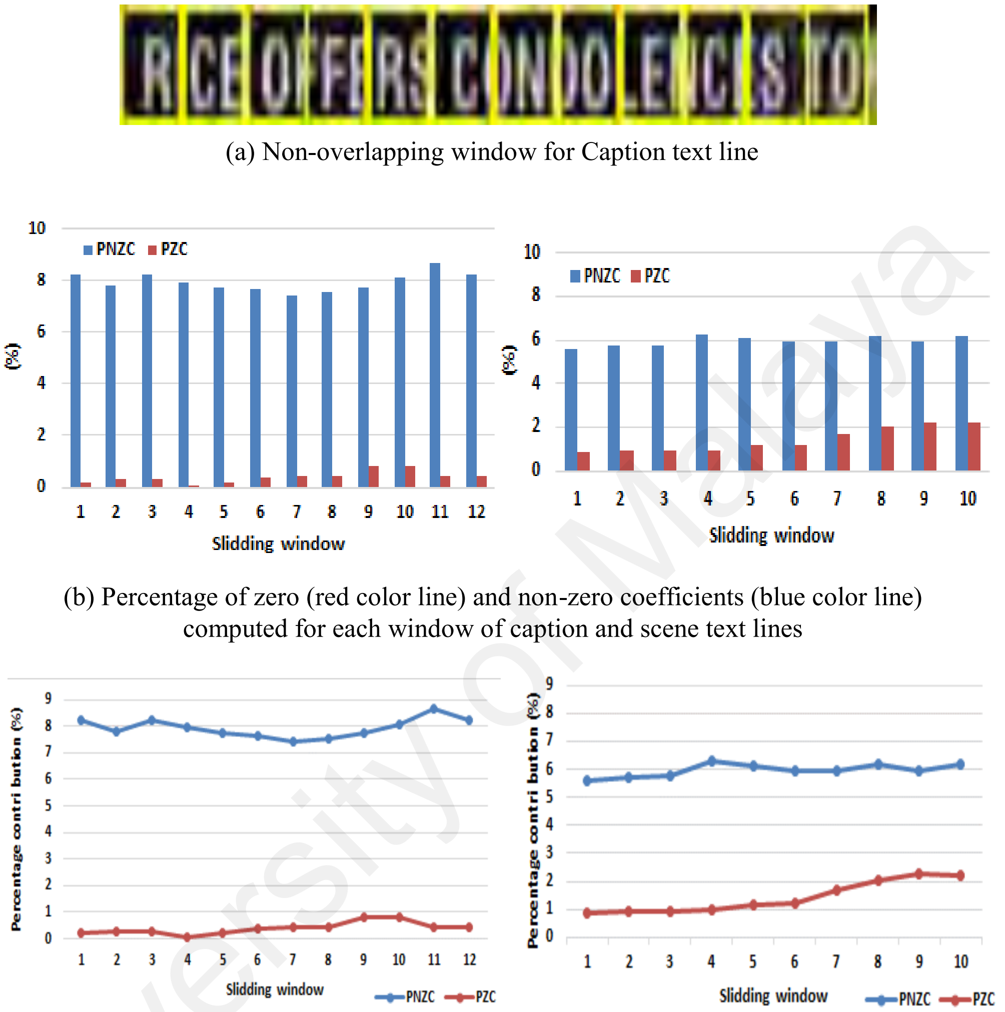
To extract such behavior of the distributions of zero and non-zero coefficients, line graphs are plotted as shown in Figure 6.2 (c) for the same values in Figure 6.2 (b), where the same observations can be visualized in the form of smoothness and non-smoothness of the lines to differentiate caption and scene texts. The line graph is mapped to image format to study linearity and smoothness of the lines.

$$PZC_W = \frac{ZC_w * 100}{ZC + NZC}, \tag{6.1}$$

where ZC_w represents the count of ZC in every sliding window w. ZC denotes the total number of zero coefficient counts in the image, and NZC refers the total number of non-zero coefficient counts.

$$PNZC_W = \frac{NZC_w * 100}{ZC + NZC}, \tag{6.2}$$

where NZC_w represents the count of NZC in every sliding window *w*.



(a) Non-overlapping window for Caption text line

(b) Percentage of zero (red color line) and non-zero coefficients (blue color line) computed for each window of caption and scene text lines

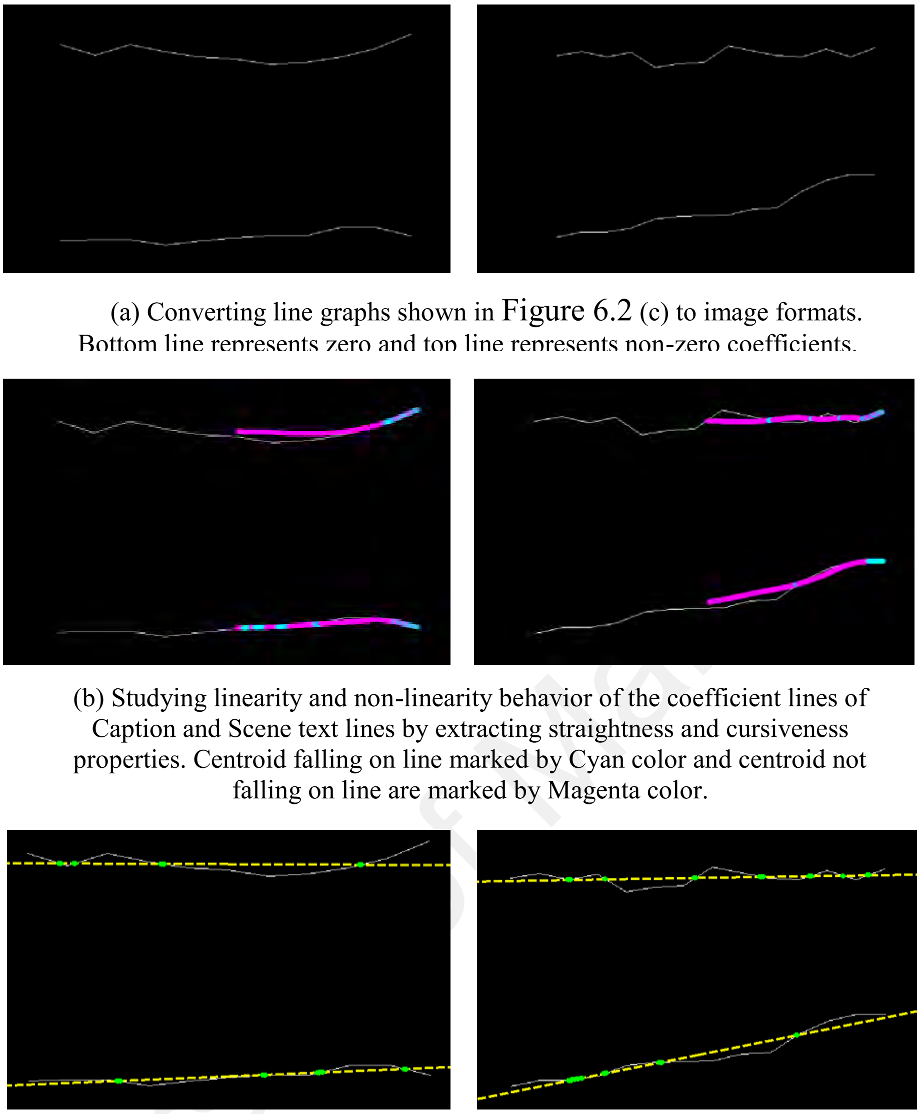
(c) Line graphs for the values in (b): Red line represents zero coefficients and blue line represents non-zero coefficients

Figure 6.2: Linear and non-linear behavior of zero and non-zero coefficients of DCT over Caption and Scene text lines.

6.2.2 Classification of Caption and Scene Text Types

To extract features for studying the behavior of coefficient lines, the lines are mapped to the spatial domain as shown in Figure 6.3 (a) for the lines shown in Figure 6.2 (c), where lines are displayed in image formats. Each line in the images is considered in Figure 6.3 (a) as the inputs for studying linearity and smoothness properties of the lines with respect to zero and non-zero coefficients of caption and scene text lines. A novel iterative method has been proposed to check whether the centroid of a line falls on itself or not. It is a fact that if the centroid falls on a line itself, the line can be considered as a straight one, or else it can be considered as a cursive line. In the first iteration, the method considers the whole line for checking whether the centroid falls on it or not. In the second iteration, the method considers the line by reducing one pixel. This process of checking centroid continues until the last pixel is reached. Further, the proposed method calculates the percentage of count (PMC) that falls on the lines as defined in equation (6.3). The process is illustrated in Figure 6.3 (b), where one can expect a larger percentage with respect to straightness for non-zero coefficient lines (top line of scene text image) than zero coefficient lines (bottom line of scene text image) line for scene texts. Similarly, the percentage of the count which represents straightness of zero coefficients (bottom line of caption text image) is lower than that of the count which represents non-zero coefficients (top line of caption text image) for caption texts. Therefore formally, Rule-1(R1) is defined for identifying tampered text as a caption as in equation (6.4).

$$PMC = \frac{count\ of\ centroid\ falling\ on\ line * 100}{total\ number\ of\ pixel\ in\ line} \tag{6.3}$$



(a) Converting line graphs shown in Figure 6.2 (c) to image formats. Bottom line represents zero and top line represents non-zero coefficients.

(b) Studying linearity and non-linearity behavior of the coefficient lines of Caption and Scene text lines by extracting straightness and cursiveness properties. Centroid falling on line marked by Cyan color and centroid not falling on line are marked by Magenta color.

(c) Studying smooth and non-smooth behavior of the coefficient lines of Caption and Scene text lines by extracting crossing points given by principal lines of text lines. Principal axis is marked in Yellow dotted color and crossing points are marked in green color.

Figure 6.3: Extracting behavior of the Caption and Scene text lines.

$$R1 = \begin{cases} 1, Scene,\ if\ PMC_{NZC} \geq PMC_{ZC} \\ 0, Caption,\ Otherwise \end{cases} \qquad (6.4),$$

where NZC denotes non-zero coefficient lines, ZC denotes zero coefficient lines, PMC denotes the percentage of a centroid falling on the respective lines. This rule (R1) helps to identify a tampered text as a caption based on the linearity of lines, which in turn helps in the classification of scene texts. Since the classification of caption and scene text is not a simple problem due to unpredictable nature of scene text, one property may not be sufficient to achieve good results. Therefore, one more novel idea has been proposed for studying smoothness and non-smoothness of the lines of caption and scene texts.

For each line in caption and scene text images, the method estimates the principal axis using the coordinates of the respective lines as shown in Figure 6.3 (c), where the yellow color dotted line is the principal axis. To understand whether the line is smooth or not, the number of crossing points (CP) done by the principal axis are counted with the lines as marked by green color in Figure 6.3 (c) for caption and scene text images. It is observed that the number of crossing points of zero coefficients and non-zero coefficients lines is almost the same for scene texts, while it is not so for captions texts. Therefore, the percentage of crossing points are calculated to define Rule-2 (R2) for classifying caption and scene text as in equation (6.5).

$$R2 = \begin{cases} 1, Scene,\ if\ |PCP_{NZC} - PCP_{ZC}| \leq 1 \\ 0, Caption,\ Otherwise \end{cases} \quad (6.5),$$

where PCP denotes the percentage of crossing points. Furthermore, the final classification combines rule-1 and rule-2 as defined in equation (6.6) and equation (6.7) for scene text and caption text, respectively.

$$\mathrm{R_{scene}} = \mathrm{\bigcup_{n=1}^{m}(R1_n, R2_n)}, \quad (6.6),$$

$$\mathrm{R_{caption}} = \prod_{n=1}^{m} (\mathrm{R1_n} = 0, \mathrm{R2_n} = 0), \quad (6.7),$$

where m denotes the total number of scene and text images.

However, the proposed method works on only full-text line but not on a word which is generally preferred to feed into OCR for recognition. In addition, although the above method works on video, the temporal feature is not exploited. Therefore, to extend the method at the word level, wavelet decomposition and temporal coherency have been explored for better text recognition in the following section.

6.3 Temporal Integration for Caption and Scene Text Types Classification at Word Level

Words segmented from video frames are the input for classifying captions and scene texts in the video in this present work. It is true that high-frequency coefficients in wavelet domain represent high contrast pixels in a text image (Liang et al., 2015a). This cue motivates to apply Haar wavelet decomposition to study high and low-frequency coefficients. According to observation, for a text image, high-frequency coefficients are positive coefficients which represent text pixels, while negative coefficients are low-frequency coefficients which represent non-text pixels (P Shivakumara, N Vinay Kumar, et al., 2014; J. Xu et al., 2014). This new observation leads to classifying positive coefficients as candidate text pixels, and negative coefficients as candidate non-text pixels. This is the main advantage of this work as it does not require any binarization methods or edge detectors for segregating non-text and text pixels as in the existing methods (S. Roy, Shivakumara, Pal, Lu, & Tan, 2016; J. Xu et al., 2014).

Since caption text has high contrast, clarity, and homogenous background, it is expected the distribution of candidate text pixels can satisfy zone pattern, namely, top, middle and bottom, which is a well-known property to find the indication of the text of any scripts. In case of scene text, since the background and foreground are unpredictable, the distribution of candidate text pixels may not generate the same zone pattern. To extract this observation, the proposed method calculates standard deviations for low-frequency values of candidate text pixels to study zone pattern. It is known that caption text stays at the same location for a few frames, while scene text moves slightly because it is a part of the background (J. Xu et al., 2014). The proposed method exploits this observation for deciding the required number of successive frames with the cues given by bins relationship. The determined temporal frames are used to find the stability of the texts. If the input video contains caption texts, the property that defines caption texts exhibits stability else the property that defines scene text exhibits stability.

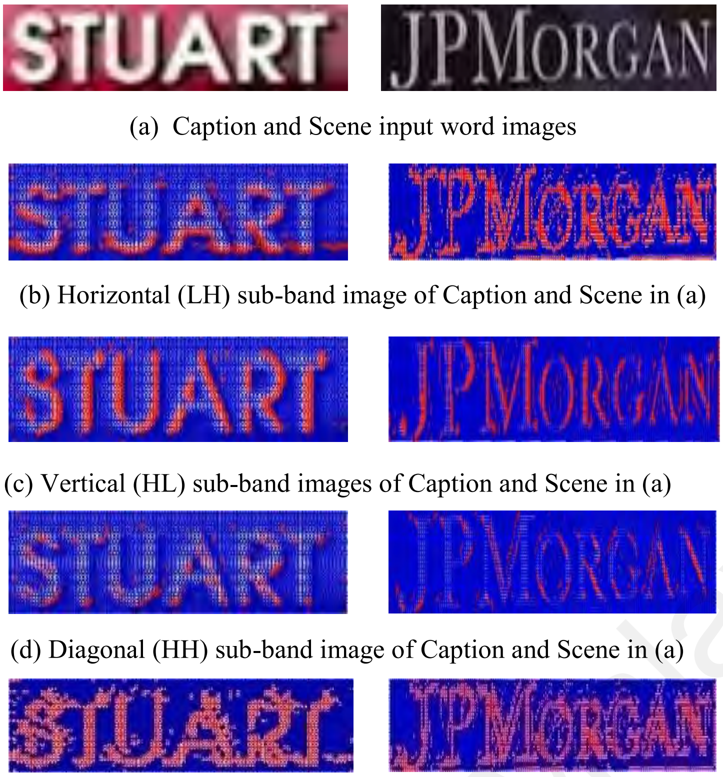
6.3.1 Wavelet Positive Coefficients for Text Candidate Detection

Candidate text and non-text pixels are obtained as defined in equation (6.8) for horizontal (LH), vertical (HL) and diagonal (HH) high-frequency bands in the previous section. It is illustrated in Figure 6.4, where for the input caption and scene words shown in Figure 6.4 (a), most of the text pixels represented by red color and non-text pixels represented by blue color pixels are classified successfully with the positive and negative coefficients in LH, HL, HH shown respectively in Figure 6.4 (b) - Figure 6.4 (d). Therefore, candidate non-text and text pixels are generated by this separation process. In order to obtain the complete structure of character components to study characteristics of caption and scene words, union operation has been performed as defined in equation (6.9) to combine candidate pixels in horizontal, vertical and diagonal frequency sub-bands as shown in Figure 6.4 (e), where better characters structure is observed compared to horizontal, vertical and diagonal. Thus, text pixels (red color) and non-text pixels (blue color) in fused results of caption and scene words shown in Figure 6.4 (e) are considered as text and non-text candidates.

$$CP_I = \begin{cases} Text, \quad W_I > 0, \quad W_I \in (LH_I, HL_I, HH_I) \\ \qquad Non-Text, \quad W_I < 0 \end{cases} \tag{6.8}$$

where CP denotes Candidate Pixels, *W* denotes wavelet coefficients in LH, HL and HH of wavelet and *I* denotes the input image:

$$\mathrm{F}_I = \bigcup(LH_I, HL_I, HH_I) \qquad (6.9).$$



(a) Caption and Scene input word images

(b) Horizontal (LH) sub-band image of Caption and Scene in (a)

(c) Vertical (HL) sub-band images of Caption and Scene in (a)

(d) Diagonal (HH) sub-band image of Caption and Scene in (a)

(e) Fused image with union of red pixels of LH, HL and HH of Caption and Scene images

Figure 6.4: Text and non-text candidates of Caption and Scene word images at level 1. Red pixels denote positive coefficients and blue pixels denote negative coefficients.

6.3.2 Cues for Caption and Scene Words from Text Candidates

One can notice from Figure 6.4 (e) that text candidates given by the previous step preserve character structures for caption word images, while for scene word images, it does not. This is valid because caption word does not have much influence of background compared to scene word (P Shivakumara, N Vinay Kumar, et al., 2014; J. Xu et al., 2014). As a result, the density of text candidates over caption word in Figure 6.4 (e) is low for first rows from top to bottom, then increases at middle rows and again decreases at the bottom rows. On the other hand, when text candidate distribution over scene word image is observed in Figure 6.4 (e), this zone pattern may not exist because of scattered text candidates.

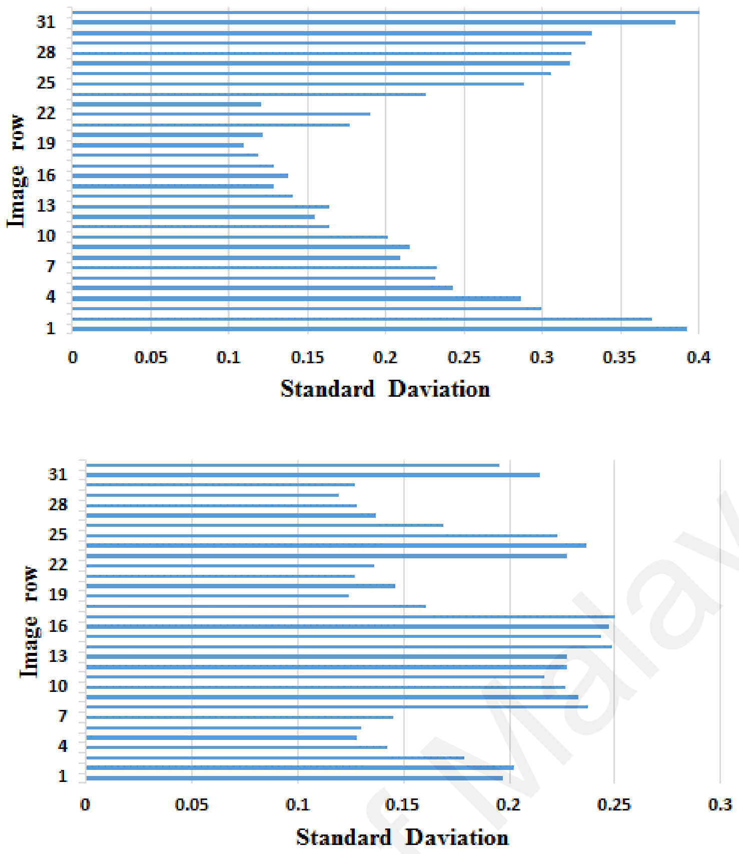
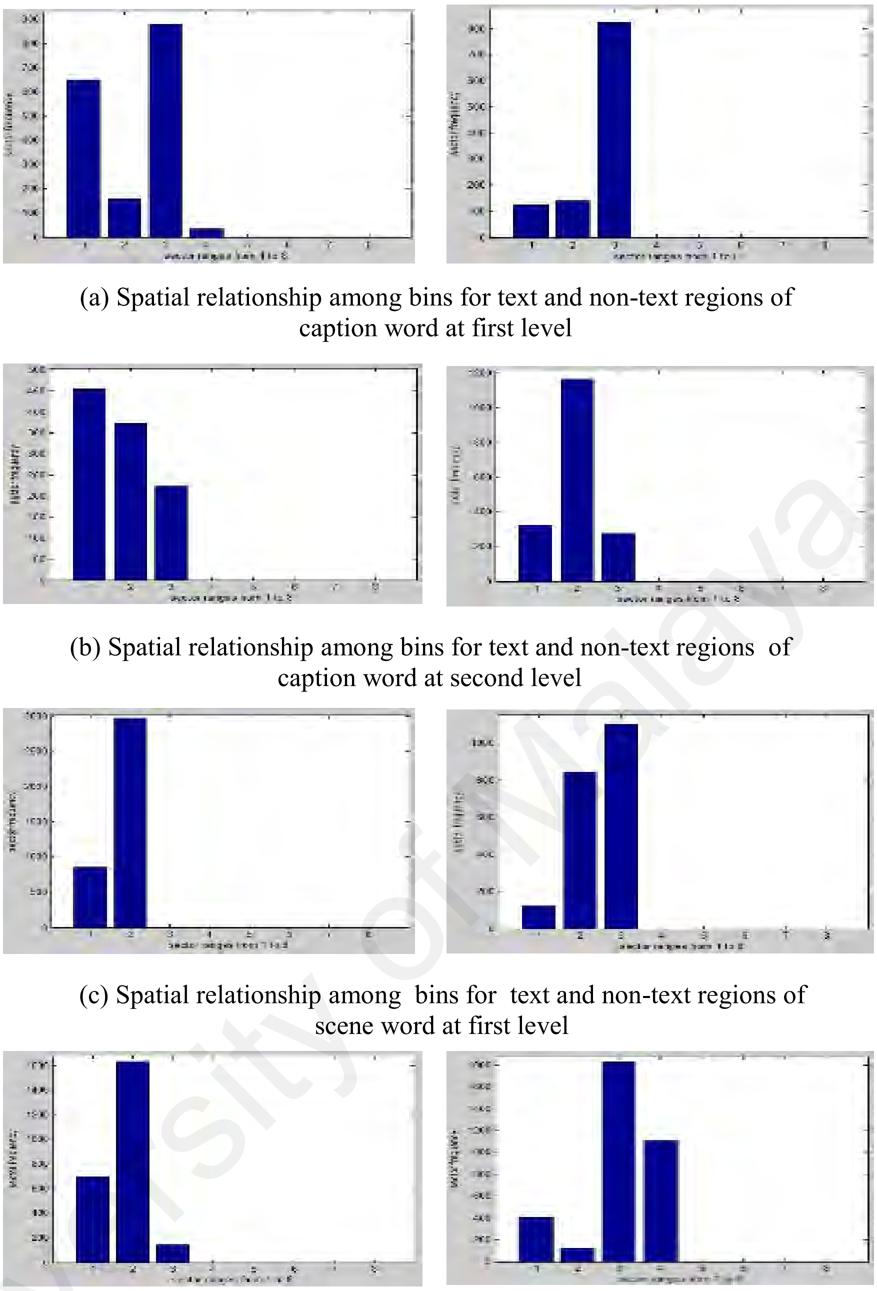


Figure 6.5: Row profile of standard deviation value of text candidates in Caption and Scene words shown in Figure 6.4 (e).

With this observation, Standard Deviation (Std) is calculated for each row of text candidate images of caption and scene texts as shown in Figure 6.5, where it is noted that caption word shows high profiles for the beginning rows, low profiles for middle rows, and high profiles for the bottom rows. For standard deviation calculation, coefficients in the low-frequency band (LL) are considered that correspond to text candidates in caption and scene word images. This is true because the middle row contains most uniform values compared to the beginning and last rows. The same zone pattern does not exist for scene word images as shown in Figure 6.5, where profile pattern does not satisfy the zone pattern as defined for caption word. Since the proposed method considers segmented words with bounding boxes for classification, different orientations of the word do not affect the classification.

To extract such distinct features, the standard deviation values have been mapped to the 8 bins formed by dividing the range of standard deviation values into 8 equal sized bins shown in Figure 6.6 (a). These bins have been studied to analyze the distribution of standard deviation values of text candidates and non-text candidates over caption and scene words. It can be seen from Figure 6.6 (a) at first level that due to fewer variations in standard deviation values, most of the standard deviation values are mapped into a few bins. Figure 6.6 (a) shows that the third bin at first level (the highest bin in the histogram) is the same for both text candidate and non-text candidate bins at first level. This is considered as Feature-1 (F_1) as defined in equation (6.10) for representing caption word at the first wavelet level. Similarly, it is expected that the total number of bins which receive standard deviation values in both text and non-text candidate sectors at the first wavelet level will be the same. This is considered as Feature-2 (F_2) as defined in equation (6.11) for representing caption word.



(a) Spatial relationship among bins for text and non-text regions of caption word at first level

(b) Spatial relationship among bins for text and non-text regions of caption word at second level

(c) Spatial relationship among bins for text and non-text regions of scene word at first level

(d) Spatial relationship among sectors for text and non-text regions of scene word at second level

Figure 6.6: Four features for representing caption and scene words.

In the same way, two more features, say Feature-3 (F_3) as defined in equation (6.12) and Feature-4 (F_4) as defined in equation (6.13) can be derived from text and non-text candidate distributions at the second level as shown in Figure 6.6 (b), where it can be seen the total number of bins which receive standard deviation values are the same for both text and non-text candidate histograms. It is also noted from Figure 6.6 (a) at first and second levels that for caption word, F_1 at the first level and F_2 at the second level represent caption words successfully, while F$_2$ at the first level and F$_1$ at the second level do not represent. Therefore, any one of the features matches out of four; this will be considered as a correct feature for representing caption word. At the same time, none of the features matches out of the four features for the input word; thus it is considered as scene word representation. This is illustrated in Figure 6.6 (c), where the highest bin position and the total number of bins of text and non-text candidate histograms at the first level do not match. The same is true for the second levels also as shown in Figure 6.6 (d). Note that according to experimental analysis, level-1 and level-2 are enough for achieving better results.

$$F_1 = \max_{l=1}\left(H_{\,p}\right) == \max_{l=1}\left(H_{\,n}\right) \qquad (6.10),$$

where l refers to level 1. $H_{\,p}$ and $H_{\,n}$ are the histograms for text and non-text candidates, respectively.

$$F_2 = \sum_{l=1}(h_P \neq 0) == \sum_{l=1}(h_n \neq 0) \qquad (6.11)$$

$$F_3 = \max_{l=2}\left(H_{\,p}\right) == \max_{l=2}\left(H_{\,n}\right) \qquad (6.12)$$

$$F_4 = \sum_{l=2}(h_P \neq 0) == \sum_{l=2}(h_n \neq 0) \qquad (6.13)$$

6.3.3 Temporal Integration for Deciding the Number of Temporal Frames

The previous section provides cues in the form of four features for classify caption and scene words. The same cues are used in this section to determine the number of frames from 25-30, captured per second. Unlike the existing methods which ignore the use of temporal coherency in video (P Shivakumara, N Vinay Kumar, et al., 2014), temporal-spatial coherency has been explored depending on the assumption that caption text stays at the same place for some successive frames, while scene text (background) has a fewer movements (J. Xu et al., 2014). To exploit this observation, in this work, non-text candidates of caption and scene words instead of text candidates are used. This is due to that non-text candidate distribution represents background, and the distribution does not differ much for both caption and scene words compared to text candidate distribution.

Therefore, the required number of temporal frames has been decided based on F4 . It is true that as long as a full text appears, there will not be a significant change in Std values in the bins. Therefore, the difference between the total number of bins that have standard deviation value in temporal frame-1 and temporal frame-2 remains the same. When gradually text disappears as temporal frames increases, changes are expected in the total number of bins. Based on this observation, the difference has been found between the total numbers of bins which received standard deviation values of the first temporal frame with other successive temporal frames, which is called as an error estimation. With the error estimation, the following conditions are set as a stopping criterion as defined in equation (6.14) for moving to the next frame. The process is illustrated in Figure 6.7 for the caption word image "LEE", where the error remains the same until frame number 13, and then there is a sudden change at frame number 14. This sudden change is considered as the stopping criterion, at the same time, the same number can be considered for deciding the types of text. The error indicates that text appears, and a sudden change in error indicates that text disappears.

$$\boldsymbol{F_n} = \begin{cases} stop, & if\ e_t < e_{t-1} and\ e_t \leq e_{t-2} \\ F_{n+1}, & if\ e_t \geq e_{t-1} and\ e_t \geq e_{t-2} \end{cases} \tag{6.14}$$

Here $\boldsymbol{e_t}$ denotes the current frame error, while $\boldsymbol{e_{t-1}}$ and $\boldsymbol{e_{t-2}}$ are the errors of previous frames.

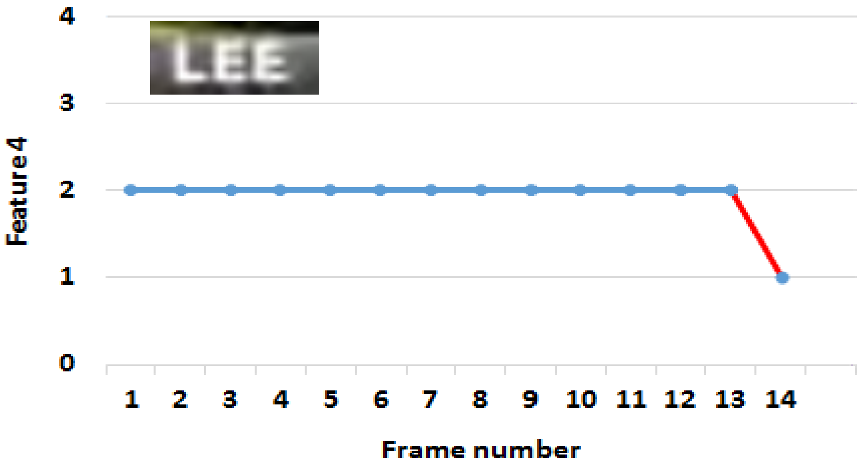
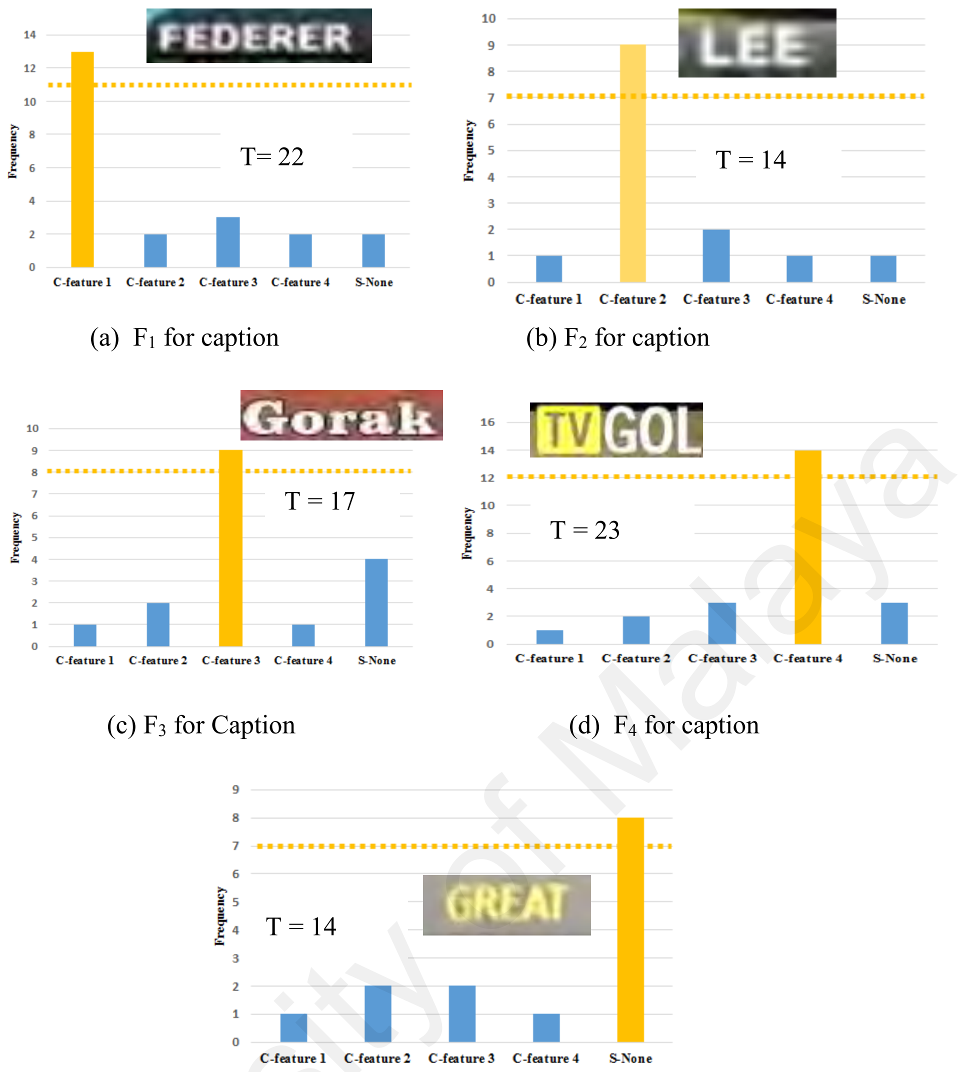


Figure 6.7: Stopping criterion for determining the number of frames. Blue line indicates for moving to next frame and red line indicates for stopping criteria which as the present error is less than immediate previous and less or equal to all previous errors.



(a) F_1 for caption (b) F_2 for caption

(c) F_3 for Caption (d) F_4 for caption

(e) Feature for scene

Figure 6.8: Procedure to choose stable features for caption and scene word classification. “Yellow” dashed line marks the base line to choose stable features and *T* is total number of temporal frames.

6.3.4 Stable Property for Caption and Scene Text Types Classification

Since the previous section provides the number of temporal frames to be used, it is good to choose stable features which correctly represent unknown input words. Section 6.3.2 gives four features for representing caption and scene text words without temporal integration. Therefore, four features are tested individually for selecting the number of temporal frames (T) given by Section 6.3.3 to choose stable features, which can represent caption or scene text words as defined in equation (6.15). For an unknown input word image, the proposed method counts the number of frames that satisfy the respective features and chooses the features which cross *T/2* frames as the stable feature of the input image. Here *T/2* is considered as the baseline for choosing stable features. If any of the four features crosses the baseline as shown in Figure 6.8 (a) - Figure 6.8 (d), where it can be seen in examples, the feature is considered as stable. Thus, it classifies the given input image as a caption word. On the other hand, if none of the four features crosses the baseline, the given word is classified as a scene word as shown in Figure 6.8 (e). Since the extracted features or cues are derived based on text and non-text candidates of caption and scene word images, the proposed features work well for different orientation, scripting language, and distortion to some extent.

$$R1 = \begin{cases} 1, Caption, & if \ \max_{T}(F_1||F_2||F_3||F_4) \geq T\ /2, \\ 0,\ Scene, & if \quad S_{None} \geq T/2 \end{cases} \qquad (6.15),$$

where *T* is the total number of temporal frames given by sub-section 6.3.3, and S_{none} denotes a scene word.

In summary, a novel method has been proposed for classifying text-type. The proposed method introduces a new idea of exploring positive and negative coefficients of wavelet decomposition for detecting text candidates. The distribution of non-text and text candidates over caption and scene word images are studied in a novel way to derive four features that give cues for the classification of caption and scene words. Moreover, temporal coherency is explored for deciding the number frames and to find stable features to classify caption and scene words correctly.

6.4 Experimental Results and Comparative Study

This section is organized as follows. Section 6.4.1 gives details of dataset and metrics for caption and scene word classification experiment. In Section 6.4.2, experiments are described for the tempered feature-based approach. Experiments on temporal integration based approach are discussed in Section 6.4.3. Comparative study of existing word classification methods and proposed method are done in Section 6.4.4.

6.4.1 Datasets and Evaluation

For evaluating the proposed approaches, standard datasets are used. Since two approaches work on different inputs, such as tempered based approach takes text line as input whereas wavelet-based approach takes word along with temporal information as an input; therefore, the database used for evaluating the performance of proposed methods are different. Similarly, the performance measure metrics also differs for both the approaches.

For evaluating tempered based classification, standard datasets which are available publicly are considered as benchmark databases, namely, ICDAR 2013 (Karatzas et al., 2013) which contains only scene text lines with large variations, ICDAR 2015 (Karatzas et al., 2015) which is slightly more complex than ICDAR 2013, and YVT (Nguyen et al., 2014). The YVT dataset has scene words with large variations in the background. The proposed caption and scene classification methods are evaluated in terms of text detection rate, classification rate and recognition rate. The above datasets contain 28, 49 and 30 videos, respectively. This totally gives 1150 frames for experimentation. It can be noticed from the above datasets that all the three databases contain only scene texts but not caption texts. Generally, video of news channels, movies, sports, etc., frames with caption texts, frames with scene texts, and frames with caption-scene texts are all common. Therefore, the database is created by collecting videos from YouTube, where different varieties of videos with different fonts, font sizes, backgrounds, and contrasts are considered. In total, 32 videos that contain from 2-3 minutes to 15-20 minutes of content are collected. This gives 350 frames with only caption texts, 180 frames with only scene texts, and 300 frames with scene+caption texts for experimentation. Thus, in total, 900 caption and 650 scene lines have been considered for evaluating the proposed classification method.

As mentioned in the beginning of the section, for wavelet decomposition based approach, benchmark databases (DB) having temporal information, namely, ICDAR 2015 video(Karatzas et al., 2015), and YVT video (Nguyen et al., 2014) have been used. In addition, Car License Plate (CLP) video (Dlagnekov & Belongie, 2005) which contains low resolution images with distortions are used for experimentation. 664, 263, and 879 video clips, respectively for ICDAR, YVT and CLP databases are considered. Since these three databases provide only scene words, Caption Text Database (CTD) has been created by collecting news video from YouTube, which includes different fonts, font sizes, backgrounds, and contrasts. In summary, scene class contains 1806 and caption class contains 450 video clips. In total, 2256 video clips are considered for comprehensive experimentation to evaluate the proposed method.

In case of tempered based approach, the proposed classification method is evaluated at two levels, namely, (1) frame level where caption and scene texts are used for classifying frames with caption texts, frames with scene texts, and frames with caption+scene texts, and (2) line level where the caption and scene text line classification is evaluated. Text detection experiment has been performed before and after classification to validate text–type classification. Detection before classification accepts all the frames in this experiment and detection after classification accepts frames of respective classes as the input for calculating text detection accuracy. The accuracies of text detection methods after classification are expected to be higher than the accuracies before classification. This is because after classification, the same method can be modified or a different method which suits the class can be used to achieve good results. In the same way, recognition experiments are conducted to validate the classification method at the line level before and after classification. For classifying frames and text types, classification rate has been used as performance measure and for text detection, recall (R), precision (P) and f-measure (F) given by several text detection methods are used as in (Liang et al., 2015a; J. Xu et al., 2014). The instructions given in (Liang et al., 2015a) are followed for counting recall, precision, and f-measures. Furthermore, for recognition, character recognition rate is used as performance measure through several binarization methods and publicly available OCR.

Different existing text detection approaches have been used before and after classification at the frame level for comparative studies. Rong et al. (Rong et al., 2014) and Yin et al. (Yin et al., 2014) work in scene image, whereas, Khare et al. (Khare, Shivakumara, & Raveendran, 2015) explores text detection for the video frame. The aforementioned text detection approach is said to be robust to fonts, font text sizes, contrasts, text types, orientations, etc. A recent method has been implemented which classifies caption and scene texts using pixel patterns of caption and scene texts, as proposed by Xu et al. (J. Xu et al., 2014), for comparative study with the results of proposed classification method at the text line level. Similarly, comparative studies are provided with different binarization methods for before and after classification with OCR at the text line level. Howe (Howe, 2013), Su et al. (Su et al., 2013), and Milayav et al. (Milyaev et al., 2013) binarize scene word in scene image, Roy et al. (S. Roy et al., 2015) binarizes text lines in the video. These methods are capable of binarizing both scene texts and video texts.

Similarly, to measure the performance of the wavelet-based method, the same metrics which are discussed above have been calculated. Text type classification rate is calculated by generating a confusion matrix. To validate the effectiveness of the classification, recognition rates are calculated before and after classification at word level of input words. Experiments before classification consider both caption and scene words together as the input, while after classification experiment considers individual caption and scene classes separately for calculating recognition rates. The recognition rate is computed using Tesseract OCR. To show the superiority of the proposed classification approach, the recent works (Bhardwaj & Pankajakshan, 2016; S. Roy et al., 2016; J. Xu et al., 2014) which use a full-text line for classification caption and scene text are implemented. In a similar way, the same binarization methods (Howe, 2013; Milyaev et al., 2013; S. Roy et al., 2015) discussed for tempered feature based classification, are considered as the baseline methods.

6.4.2 Experiments on Tempered Feature Approach at Text Line Level

The tempered based classification method has been used for classifying frames based on the caption and scene. A few results of the proposed approach is shown in Figure 6.9. The accuracy of the proposed classification at the frame level has been tabulated in Table 6.1 in the form of the confusion matrix. To analyze the contribution of each rule for classification, confusion matrices are computed for individual rules and the combined rule as reported in Table 6.1. Table 6.1 shows that each rule contributes significantly to achieve better results. Table 6.1 shows that the combined rule scores better results than individual rules. Text detection results have been given in Table 6.2. It is found from Table 6.2 that text detection accuracies are improved significantly after classification. Therefore, to enhance the performances of text detection methods, text classification is necessary for the video.

Table 6.1: Confusion matrix of the proposed method using centroid features at text line level.

Features	Confusion matrices		
	Types	Scene	Caption
Centroid	Scene	60.52	39.48
	Caption	42.3	57.7
Crossing	Scene	63.81	36.19
	Caption	38.4	61.6
Proposed	Scene	68.66	31.34
	Caption	28.62	71.38

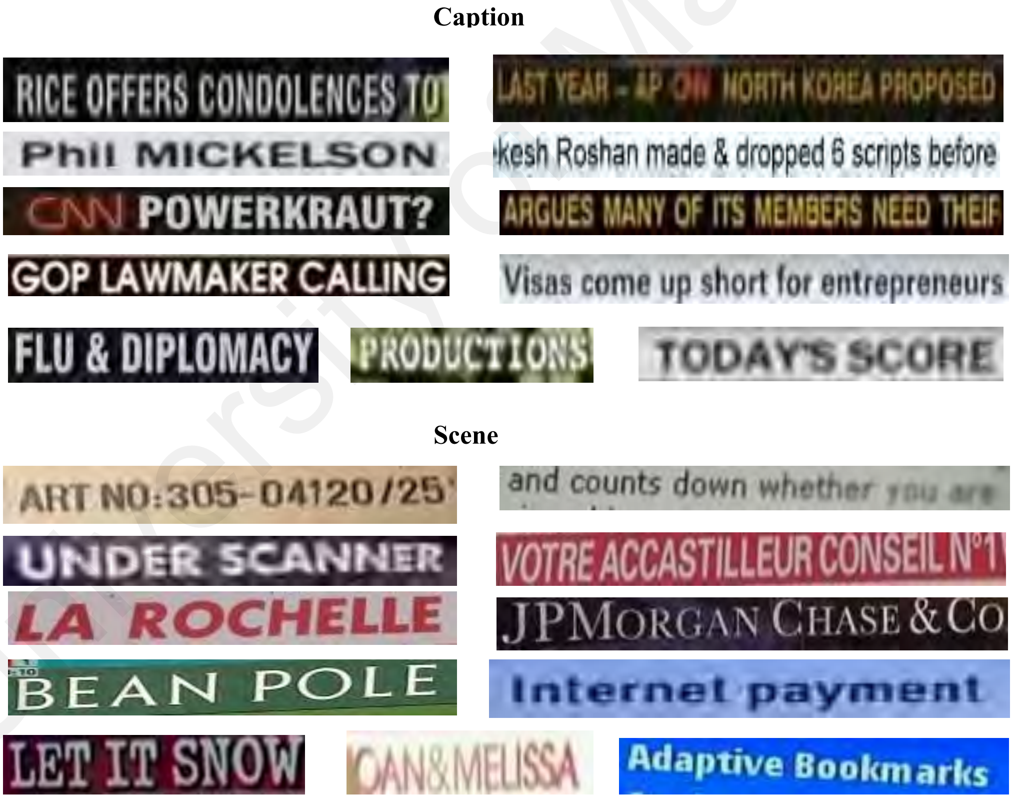

Caption
RICE OFFERS CONDOLENCES TO
LAST YEAR - AP ON NORTH KOREA PROPOSED
Phil MICKELSON
kesh Roshan made & dropped 6 scripts before
CNN POWERKRAUT?
ARGUES MANY OF ITS MEMBERS NEED THEIR
GOP LAWMAKER CALLING
Visas come up short for entrepreneurs
FLU & DIPLOMACY
PRODUCTIONS
TODAY'S SCORE
Scene
ART NO:305-04120/25
and counts down whether you are
UNDER SCANNER
VOTRE ACCASTILLEUR CONSEIL N°1
LA ROCHELLE
JPMORGAN CHASE & CO
BEAN POLE
Internet payment
LET IT SNOW
DAN&MELISSA
Adaptive Bookmarks

Figure 6.9: Samples of successful classification results of the proposed method for the Caption, Scene lines.

6.4.3 Experiments on Temporal Integration Approach at Word Level

The main steps of the wavelet-based method are 1) introducing four features (F_1-F_4) for representing caption and scene words, 2) deciding the temporal frames (T) numbers, and

3) finding stable features with temporal frames for classification. To analyze the effectiveness of the above-mentioned steps and the contribution of each feature, classification rates are calculated through a confusion matrix for scene class (1806 video clips) and caption class (450 video clips). The quantitative results are reported in Table 6.2, where one can note that for determining the number of frames (T), Feature-4 is the best compared to the other features. It is observed from Table 6.2 that the classification rates of F_1 to F_4 using keyframes, which are selected using predefined software (without temporal frames), are almost same. This shows that the four features contribute equally to classification. Table 6.2 shows that the classification rates of the four features using temporal frames are higher than those without temporal frames. This indicates that the use of temporal frames is the best to achieve better classification rates.

Table 6.2: Classification rates in (%) for evaluating intermediate steps of the proposed method.

Steps	F_1	F_2	F_3	F_4
T (Number of temporal frames)	63.02	62.3	61.06	68.83
Key frame (without temporal frames)	55.24	57.26	57.01	54.81
With Temporal	60.37	59.6	60.47	58.52

6.4.4 Comparative Study and Discussion

In case of tempered based classification, sample images of the proposed and existing method (J. Xu et al., 2014) are shown in Figure 6.10 for caption and scene texts at the line level. From Figure 6.10, it is seen that the proposed classification approach classifies different text types successfully, and at the same time, the existing method fails to classify them. It can be validated by the quantitative results given in Table 6.3 that the proposed approach is better than existing works. The reason is that the proposed method uses frequency coefficients while the latter relies on pixels in the spatial domain.

Table 6.3: Performance of the proposed and existing methods for caption and scene text classification at text line level (in %).

Types	Proposed method		(J. Xu et al., 2014)	
	Scene	Caption	Scene	Caption
Scene	68.66	31.34	65.69	34,31
Caption	28.62	71.38	32.62	67.38

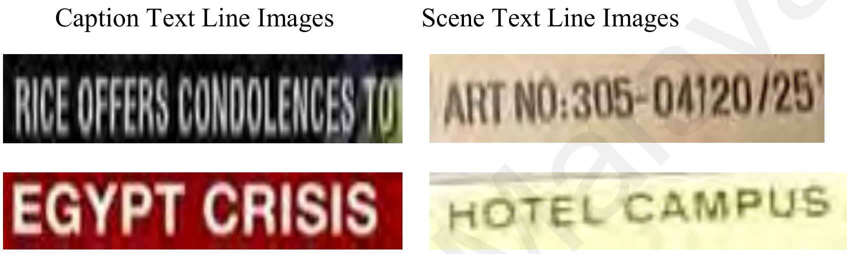


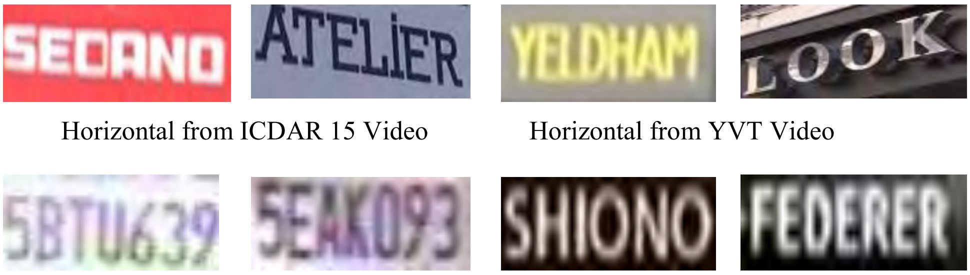
Figure 6.10: Examples of qualitative results. The proposed method classifies the above text line images successfully while the existing methods misclassify.

To show the usefulness of the classification at the text line level, recognition results of different binarization methods before and after classification are tabulated in Table 6.4. From this table, it is seen that the recognition rate has been increased after classification than before classification. The same methods have been used for both before and after classification tuning the parameters based on samples of each class by considering advantage of classification step. The key parameter chosen from each method has been tuned such as the threshold value for the Canny image (Howe, 2013), the threshold value for the Bayesian classifier (S. Roy et al., 2015), and window size for binarization algorithms (Milyaev et al., 2013). Su et al. (Su et al., 2013) provides exe file, thus so no tuning has been applied. As a result, improvements can be observed in the recognition results after classification. To achieve good results in different types of images like plain document images, one can modify the method or use a different method since texts have been classified.

Table 6.4: Character recognition rates of different binarization methods before and after classification at text line level (in %).

Methods	Before classification		After classification			
	Scene and caption		Scene		Caption	
(Howe, 2013)	51.33	0.4	56.31	0.4	51.78	0.2
(S. Roy et al., 2015)	40.16	0.05	42.48	0.2	47.29	0.03
(Su et al., 2013)	48.9		55.2		49.6	
(Milyaev et al., 2013)	42.67	10	49.58	10	45.7	8

In case of the wavelet-based method, sample images of the proposed and existing approaches (Bhardwaj & Pankajakshan, 2016; S. Roy et al., 2016; J. Xu et al., 2014) are shown in Figure 6.11 for different databases. The proposed method classifies different word type images successfully including multi-orientation texts, while the existing method is unsuccessful. The reason for the poor classification of the existing method (J. Xu et al., 2014) is that it obtains text candidates using stroke width and gradient information. In addition, the existing method requires full-text lines for achieving better results Moreover, the proposed method works well at word level without binarization. The quantitative result obtained from the proposed and existing approaches is given in Table 6.5. From the table, it is noted that the performance of the proposed approach is better than the existing method. From Table 6.5, it is seen that the proposed and state-of-the-art methods give better results for ICDAR + CTD, slightly low results for YVT + CTD and the lowest for CLP + CTP. It is evident that YVT data contains non-horizontal word and CLP data is more complex than the other two databases because it includes blur and distorted word images.



Horizontal from ICDAR 15 Video Horizontal from YVT Video

Sample words from CLP database Sample words from CTD

Figure 6.11: Examples of qualitative results. The proposed method classifies the above word images successfully while the existing methods misclassify them.

Table 6.5: Confusion matrix of the proposed and existing methods on different databases.

DB	ICDAR 2015+CTD			YVT+CTD		CLP + CTD	
Methods	Type	Caption	Scene	Caption	Scene	Caption	Scene
Proposed	Caption	71.68	28.32	71.68	28.32	71.68	28.32
	Scene	27.51	72.49	35.64	64.36	39.41	60.59
(J. Xu et al., 2014)	Caption	57.0	43.0	57.0	43.0	57.0	43.0
	Scene	46.54	53.46	49.3	50.7	52.48	47.52
(S. Roy et al., 2016)	Caption	65.72	34.28	65.72	34.28	65.72	34.28
	Scene	33.2	66.8	32.86	67.14	37.31	62.69
(Bhardwaj & Pankajakshan, 2016)	Caption	70.02	29.8	70.02	29.8	70.02	29.8
	Scene	31.68	68.32	35.9	64.1	36.6	63.4

To validate the effectiveness of the wavelet-based classification, experiments are conducted on recognition before and after classification similar to DCT based classification. The results of the binarization methods are listed in Table 6.6. It is noted from Table 6.6 that performances of the binarization techniques are better in case of after classification than before classification. This is because of the advantage of classification, where tuning or modification of the method according to the complexity of the problem can be done to achieve better results. For example, for the methods listed in Table 6.6, parameters such as the threshold value for Canny edge image in case of (Howe, 2013), the threshold value for Bayesian classifier in case of (S. Roy et al., 2015), and the window size for the method in (Milyaev et al., 2013) are tuned. Note that the method in (Su et al., 2013) provides only exe file; therefore there are no parameter values listed in Table 6.6. Table

6.6 shows that with a few modifications to the parameter after classification, the methods show better results compared to before classification. Therefore, it can be inferred that classification is useful to achieve better results.

Table 6.6: Recognition rates (RR) at word level of the different binarization methods before and after classification (in %). 'P' denotes a parameter.

Methods	Before classification		After classification			
	Caption + Scene		Caption		Scene	
	RR	P	RR	P	RR	P
(Howe, 2013)	52.7	0.4	54.38	0.2	53.63	0.4
(S. Roy et al., 2015)	42.75	0.05	49.25	0.03	48.32	0.2
(Su et al., 2013)	52.1	-	51.4	-	53.82	-
(Milyaev et al., 2013)	45.3	10	53.3	8	51.6	10

6.5 Summary

Overall, in this chapter, two video text type classification approaches based on DCT and wavelet are presented for better text recognition. The first approach explores DCT for identifying tampering information from the text line, whereas the second approach explores temporal coherency in addition to wavelet coefficient from the word. Although both the approaches take different input for classification, they exploit the same fact that caption text is edited/superimposed, which results in artificially created text comparing to other frame contents.

CHAPTER 7: MULTI-TYPE-ORIENTED VIDEO TEXT RECOGNITION

7.1 Background

In the previous chapter, classification of caption and scene text in the video is presented. This chapter studies advantages and disadvantage of recognizing text in the video with binarization and without binarization. For recognition through binarization, the work explores Bayesian classifier-based method and fusion of machine learning approaches mainly, Support Vector Machine (SVM) and Hidden Markov Model (HMM).

7.2 Binarization based Text Recognition

In this work, words are segmented from text lines for recognition because of the availability of several methods in the literature which work well at the word level regardless of the type of text, orientation, etc. (N. Sharma, Pal, et al., 2012). Segmented words rather than text lines have been used for recognition because most of the recognition methods calculate recognition rate at the word level and because words are used as units for navigation, tracking, and surveillance. Since words in the video have low resolution, color, wavelet and gradient sub-bands have been integrated to increase low resolution text information by widening the difference between background and text region. RGB separates background and text information according to (Palaiahnakote Shivakumara, Trung Quy Phan, et al., 2010), because RGB offers more information from the three channels which allow to choosing the most prominent edge pixels. It is also shown that wavelet sub-bands (LH, HL, HH) are useful for classification between text and non-text pixels (Palaiahnakote Shivakumara et al., 2014) as they provide spatial information. Here, wavelet with high-frequency sub-bands is explored for binarizing video text. In the same way, gradient sub-bands (horizontal, vertical and diagonal) can also be used for classification between non-text and text pixels (S. Roy et al., 2012).

The above observations motivate to propose to integrate RGB, wavelet, and Gradient. This is valid because for each one sub-band the value of a text pixel is different. The fusion of sub-bands explores this clue for increasing the information of text. From integrated results, it is observed that the pixels with high probability correspond to text whereas non-text pixels have low probability value. Besides, Bayesian uses a probabilistic approach to predict the most likely edges of text. This clue leads to propose a Bayesian classifier for binarization as it is motivated by the work presented in (Shivakumara et al., 2012) where it has been used for text detection. The enhanced images from different domains help to calculate a priori probabilities and conditional probabilities for the non-text and text pixels. Then finally, the Bayesian classifier has been proposed to predict actual text pixel for the word image in the video. To recognize the word, an OCR engine (Tesseract) has been used which is available publicly. The flow diagram of the method can be seen in Figure 7.1.

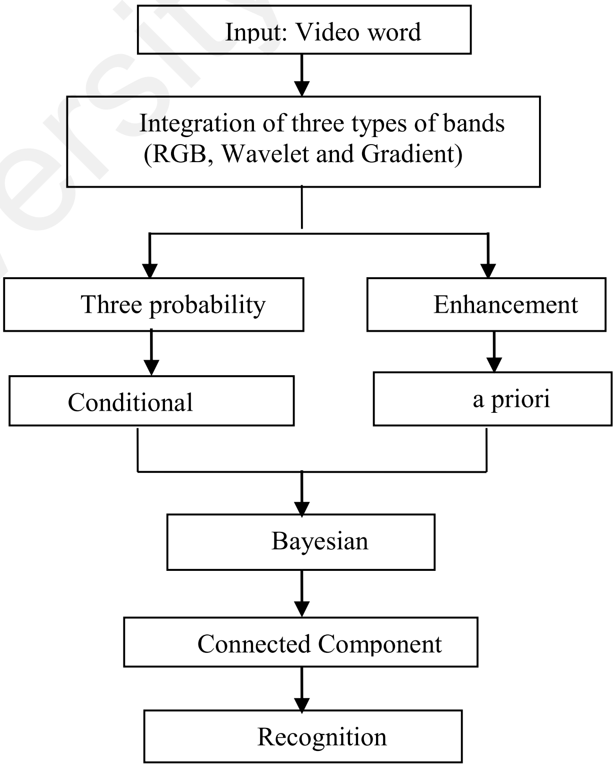
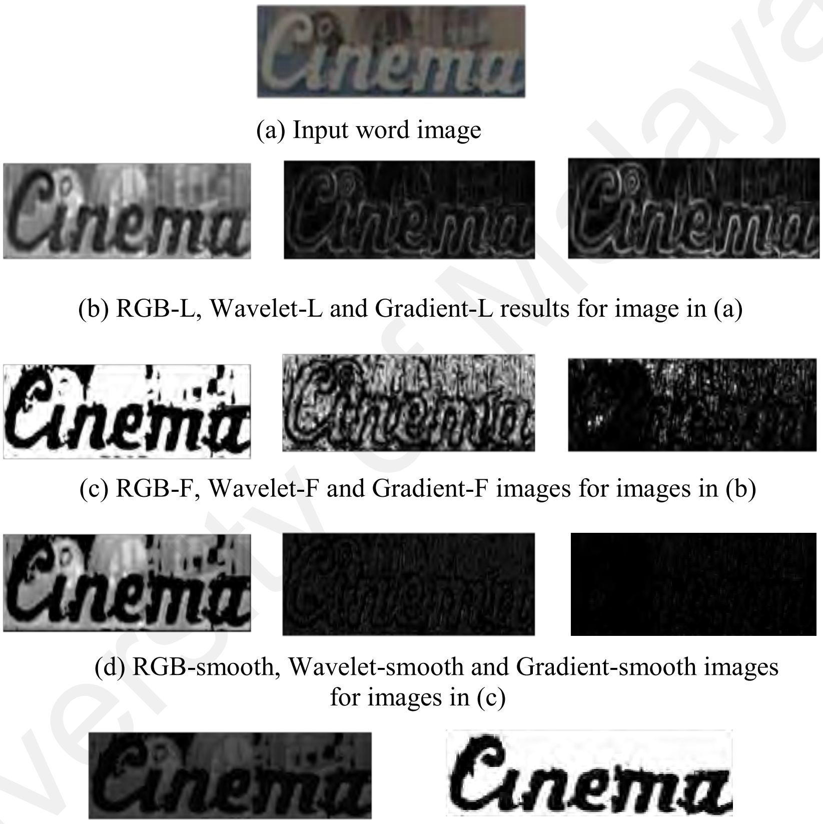


Figure 7.1: Flow diagram of the proposed method.

7.2.1 Text Enhancement Integrating Color, Wavelet, and Gradient

It has been observed from work proposed in (S. Roy et al., 2012) where fusion concept has been used for video text binarization successfully for horizontal text. This motivates to propose a new idea that fuses not only wavelet and gradient as in (S. Roy et al., 2012) but also color in a novel way for binarizing word of any orientation. For each word image shown in Figure 7.2 (a) where it can be seen that text is embedded in a complex background with some orientation, the method decomposes the input image into R, G, B sub-bands in the color domain, LH (Horizontal), HL (Vertical), HH (Diagonal) in the wavelet domain and Horizontal, Vertical, Diagonal in the gradient domain. After that, three sub-bands have been integrated linearly (L) to generate three fused images, namely, RGB-L, Wavelet-L and Gradient-L as shown in Figure 7.2 (b). Figure 7.2 shows fine details of the image. The linear integration is a simple operation which integrates three pixels values from three sub-band images for one target pixel. This results in an enhanced image. Due to low contrast, sometimes noise pixels are also fused with text in this linear combination process. Therefore, to remove noisy pixels from the enhanced image, filtering approach has been proposed.

The method performs histogram analysis for each sliding window over the enhanced image obtained from the linear operation. The height and width of the window are the same as the height of the word. This is possible because the input word segmented by the segmentation method gives height and width of the word image. The method selects the top three dominant values from the histogram of each sliding window as text pixels. Noisy pixels are thus filtered out due to their low values. With this operation, three filtered images have been obtained from three domains, namely RGB-F, Wavelet-F and Gradient-F as shown Figure 7.2 (c) where one can see the enhancement of text pixels compared to the background. Due to background contrast variations in video, there might be non-text pixels that exhibit high contrast values as text pixels. To overcome this problem, for each pixel in the filtered image, the method again performs a window operation to choose median values as text pixels, eliminating high contrast non-text pixels because, as median values, they are neither high and nor low. With this, three smoothed images have been obtained for the above three domains, respectively, RGB-Smooth, Wavelet-Smooth, and Gradient-Smooth as shown in Figure 7.2 (d) where it connects missing text pixel and eliminates noisy pixels.



(a) Input word image

(b) RGB-L, Wavelet-L and Gradient-L results for image in (a)

(c) RGB-F, Wavelet-F and Gradient-F images for images in (b)

(d) RGB-smooth, Wavelet-smooth and Gradient-smooth images for images in (c)

(e) Integrated image

(f) Binarization result

Figure 7.2: Text enhancement by integrating sub-bands.

Finally, the three smoothed images are combined using weight. The weight is determined based on 100 samples chosen randomly from databases with the help of a classifier. A scheme proposed in (M. Li, Zhang, & Mao, 2008) has been used for determining weights. Variance is computed in a local region of the target pixel. It helps in choosing the high-frequency coefficients. The operation is expressed as follows:

$$\sigma_{RGB}(I) = \frac{1}{M \times N} \sum_{c=-\frac{M}{2}}^{c=\frac{M}{2}} \sum_{d=-\frac{N}{2}}^{d=\frac{N}{2}} I\,(s+c, t+d)$$

Where the size of the neighbor or local area of the target pixel is $M \times N$, and, σ_{RGB} denotes the variance of the coefficients placed at (s,t) location in the window of $M \times N$, respectively. Since the square window has been considered as mentioned above, M and N will have the same value. Then, the fusion scheme can be defined as follows:

$$D_F = \begin{cases} D_{RGB,} \ \ \sigma_{RGB}\ (I) \geq \sigma_W(I) \geq \sigma_G\,(I) \\ D_{W,} \ \ \sigma_W\,(I) \geq \sigma_{RGB}(I) \geq \sigma_G\ \ (I) \\ D_{G,} \ \ \sigma_G\ \ (I) \geq \sigma_{RGB}\ (I) \geq \sigma_G\,(I) \end{cases}$$

Here, $D_{RGB,}\ D_{W,}\ D_{G,}$ refers to RGB-Smooth, Wavelet-Smooth and Gradient-Smooth. As a result, an enhanced image (integrated image) has been obtained as shown in Figure 7.2 (e) where a clear difference can be seen between text and non-text pixels. The logical steps of the enhancement process are shown in Figure 7.3 to illustrate the flow of the work.

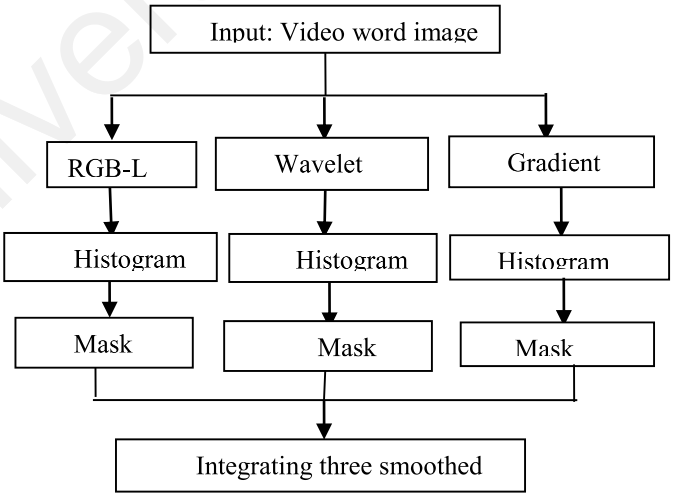
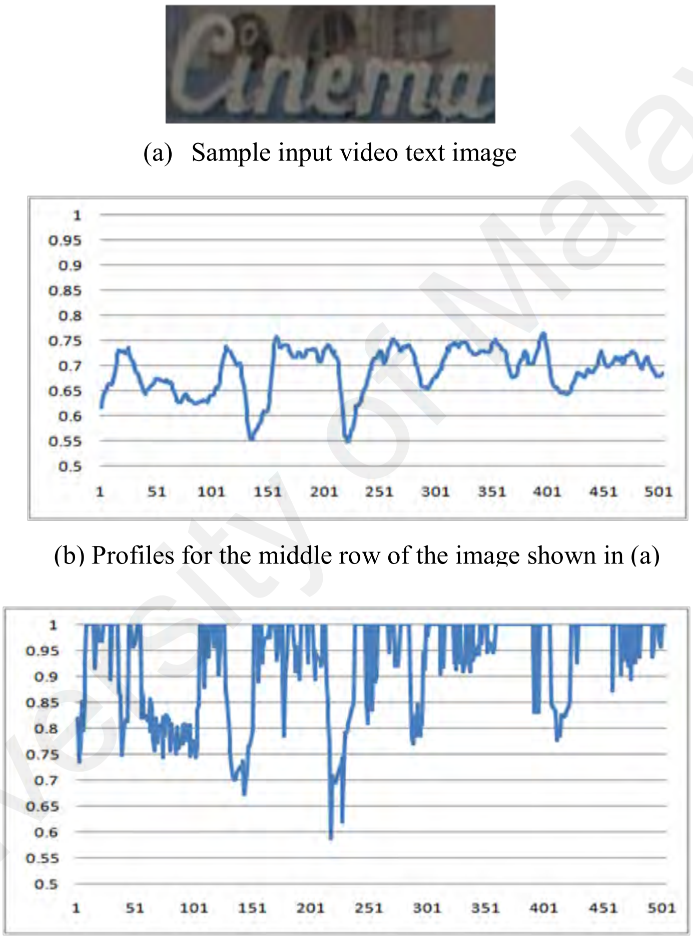


Figure 7.3: Integrating three domains to obtain enhanced image.

The enhancement step is illustrated in Figure 7.4 to show that the integration of sub-bands enhances text pixels compared to non-text pixels. In Figure 7.4, (a) is an input image, (b) is a middle-row profile of the input image, and (c) profile for the integrated image given by enhancement algorithm. From Figure 7.4 (b) - (c), it is seen that high peaks represent text pixel and low peaks represent non-text pixels in (c) compared to the peaks in (b).



(a) Sample input video text image

(b) Profiles for the middle row of the image shown in (a)

(c) Profile for the same middle row after enhancement

Figure 7.4: Illustrations for the enhancement by liner operation.

7.2.2 Bayesian Classifier for Binarization

The probability of text pixels on the fused image, obtained from fusing RGB-smooth, wavelet-smooth and gradient-smooth is high compared to the background. This cue leads to propose a Bayesian classifier since it is a two class problem. Therefore, a simple probability has been proposed for classifying non-text and text: If a pixel has high value (towards 1) at the same location of the three images (in RGB, Wavelet and Gradient images), then the probability of that pixel is considered as high to be a text pixel. In the same way, if the value of a pixel is low (towards 0), then the probability of that pixel is high to be a background or non-text pixel. This observation motivated to propose a Bayesian classifier for binarization in this work. The proposed method calculates the conditional probability for text pixel given text class as defined in equation (7.1) and (7.2). Here the conditional probability is defined by P(f(x,y)) of a pixel (x, y) for a given Text Class (TC). RGB-Smooth, Wavelet-Smooth and Gradient-Smooth are defined as P1(f(x,y)), P2(f(x,y)) and P3(f(x,y)), respectively. f(x,y) corresponds to respective filtered images, namely, RGB-F, Wavelet-F and Gradient-F as mentioned earlier. Before calculating the conditional probability, normalization has been done on the values in the smoothed images to the range of 0 and 1 by dividing its maximum value. Therefore, when the three smoothed images contain high values, say 1 then that pixel is classified as text pixel. Similarly for P(f(x,y)|NTC) as given in equation (7.2), where NTC denotes Non-Text group, N1(f((x,y)), N2(f(x,y)) and N3(f(x,y)) are the complement of P1(f(x,y)) P2(f(x,y)) and P3(f(x,y)), respectively.

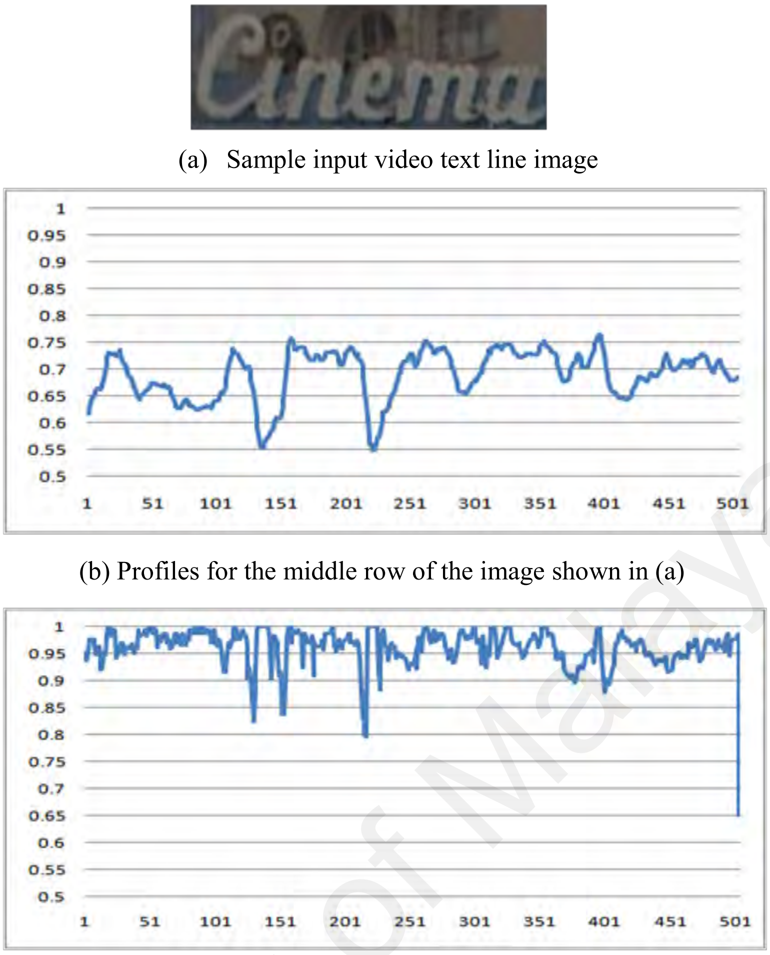
$$P(f(x,y)|TC) = \frac{\big(P1(f(x,y)) + P2(f(x,y)) + P3(f(x,y))\big)}{3} \tag{7.1}$$

$$P(f(x,y)|NTC) = \frac{\big(N1(f(x,y)) + N2(f(x,y)) + N3(f(x,y))\big)}{3} \tag{7.2}$$

In order to find a priori probability without knowing the behaviour of the dataset, k-means algorithm has been applied by integrating the procedures described in Section 7.2.1 on the enhanced image obtained to separate background and text pixels. Since the integrated image contains low values for background pixels compared to text pixels and it is a two-class problem, k is set to 2 to separate non-text pixels from text pixels. Since the k-means is unsupervised, the cluster having higher average is considered as a text otherwise, it is a non-text. In addition, to avoid misclassification of text pixels, (S. Roy et al., 2012) based method has been proposed on a fusion concept to obtain one more enhanced image. The fusion concept for enhancement is illustrated in Figure 7.5. Figure 7.5 shows the profiles drawn at the middle row of the input image (a), before and after enhancement. It is observed that the profile obtained 'after enhancement' gives high peaks with high scale compared to the profile of 'before enhancement'. This shows enhancement by fusion by segregating non-text pixels and text pixels. Hence, k-means has been applied on an enhanced image obtained by fusion. As a result, two text clusters and two non-text clusters have been obtained. From these two text clusters, the number of non-text pixels and text pixels are calculated for the whole image. The count of background and text pixels are assumed as a priori probability. The priori probability of text is denoted as P(CTC) and the priori probability of non-text is as P(NCTC). Substitution has been done in the Bayesian formula as defined in equation (7.3) to obtain posterior probability matrix. Finally, the bi-level image (B (x,y)) is generated according to equation (7.4) on posterior probability matrix. The effect of Bayesian classifier can be seen in Figure 7.2 (f) where one can notice that the method classify text and background clearly for different orientation text.

$$P\big(TC\big|f(x,y)\big) = \frac{P(f(x,y)|TC) * P(CTC)}{P(f(x,y)|TC) * P(CTC) + P(f(x,y)|NTC) * P(CNTC)} \tag{7.3}$$

$$B(x,y) = \begin{cases} 1, & P\big(TC\big|f(x,y)\big) \geq 0.05 \\ 0, & Otherwise \end{cases} \tag{7.4}$$



(a) Sample input video text line image

(b) Profiles for the middle row of the image shown in (a)

(c) Profile for the same middle row after fusion

Figure 7.5: Illustrations for the enhancement by fusion.

Sometimes, the above process may output characters with disconnections as the resolution of the video is often low. Therefore, connected component based analysis has been applied to smooth the contours. It is observed from the results shown in Figure 7.2 (f) that the Bayesian classifier retains the characters shape without any disconnections. If any disconnections appear in the output, the method identifies them by applying a mutual nearest neighbour technique on endpoints. Before testing mutual nearest neighbour, the method generates the contour of the characters using the Canny image obtained from the binary image. The mutual nearest neighbour principle is formulated as: if P_1 is close toP_2, then P_2 should be close to P_1, where P_1 and P_2 are the two endpoints. The reason for doing such operation is that Canny generates dense edge for text images, but it introduces a lot of noisy pixels in the background at the same time. Next, the comparison is performed between the identified disconnection area and Canny image at the same position to retrieve the missing text pixels. Finally, flood fill algorithm has been applied to fill the characters before sending to the OCR module.

However, due to the inherent limitation of Tesseract OCR (Smith, 2007), it is not robust to different fonts and orientations. Moreover, obtaining clear binary text image where character component shapes are preserved, is not always possible before feeding it into OCR in case of the complex video image. Therefore, a learning-based method has been explored to improve the recognition rate even for different oriented characters in the video.

7.3 Classifier-based Text Recognition

The main issue with classifier-based recognition is defining window size for arbitrarily-oriented characters in multi-type images. To overcome such issue, automatic window size detection has been proposed based on the fact that directions of most pixels contribute towards character height, which helps to fix correct windows according to sizes and orientations of characters. Further, the integration of strength of different types of features, namely, statistical features which extract geometrical properties, texture features which extract appearances property, run-length smearing which extracts intra and inter symmetry of character components, and contour wavelet domain which is invariant to scaling, multi-fonts or multi-sizes, helps to achieve better results for text in multi-type images.

7.3.1 Automatic Window Size Detection

For each word, the height of word is considered as the width of the initial window, which results in a square window as depicted in Figure 7.6 (a) and (b), where the initial square window is observed for the input word. In general, since the text detection method fixes a bounding box for the whole word by covering extra background information, the square window covers more than one character. As a result, the defined square window does not cover only one character. Therefore, to determine a correct window, wavelet high-frequency sub-bands and a fused band have been explored. The reason to propose Haar wavelet high-frequency sub-bands is that wavelet decomposition is good for classification of non-text and text (Huiping Li et al., 2000). For the square window, high-frequency sub-bands, namely, Horizontal, Vertical, and Diagonal have been obtained as shown in Figure 7.6 (b). The proposed method performs OR operation to fuse three high-frequency sub-bands as shown in Figure 7.6 (b) with label fused. Then k-means clustering with k=2 has been applied on three sub-bands and fused window to obtain respective text clusters as shown in Figure 7.6 (c). The cluster having the highest mean is represented as text. Since text pixels have high contrast values than its background (L. Wu et al., 2015), the pixels with high contrast values forms text cluster. This result outputs the structures of character components as shown in Figure 7.6 (c).

It is true that most pixel direction contributes towards the height of character components. If the angle of such a character component is estimated, then the same angle gives almost the angle of character direction. For example, if a character is in the horizontal direction, it gives almost 90 degrees. Inspired by this observation, the angle for the fused result and text cluster of high-frequency sub-bands is calculated by passing coordinates of pixels to Principal Component Analysis (PCA) as shown in Figure 7.6 (c), where the principal axis is drawn for high frequency sub-bands and the fused window. The reason to choose PCA is that it does not require the full shape of the character to find its direction (R. C. Gonzalez & Woods, 2002) as can be noticed from Figure 7.6 (c). It is true that PCA is popular for dimensionality reduction rather than angle estimation. However, the property that PCA outputs principal axis is explored for objects when two-dimensional data, such as X and Y coordinates of "0" and "1" pixels is fed in the image.

Therefore, the principal axis can be drawn using the first Eigenvector of PCA to estimate the correct angle of the character components. It is also true that since the initial square window covers one or more characters according to observation as shown in Figure 7.6 (a). As a result, the arbitrary orientation of a text does not reflect in the content of the square window. It is evident from Figure 7.6 (a) - (b), where the content of the initial square window looks horizontal with a bit tilt. Therefore, angles of vertical, and diagonal windows are ignored and the angles of horizontal and fused windows are considered for experimentation.

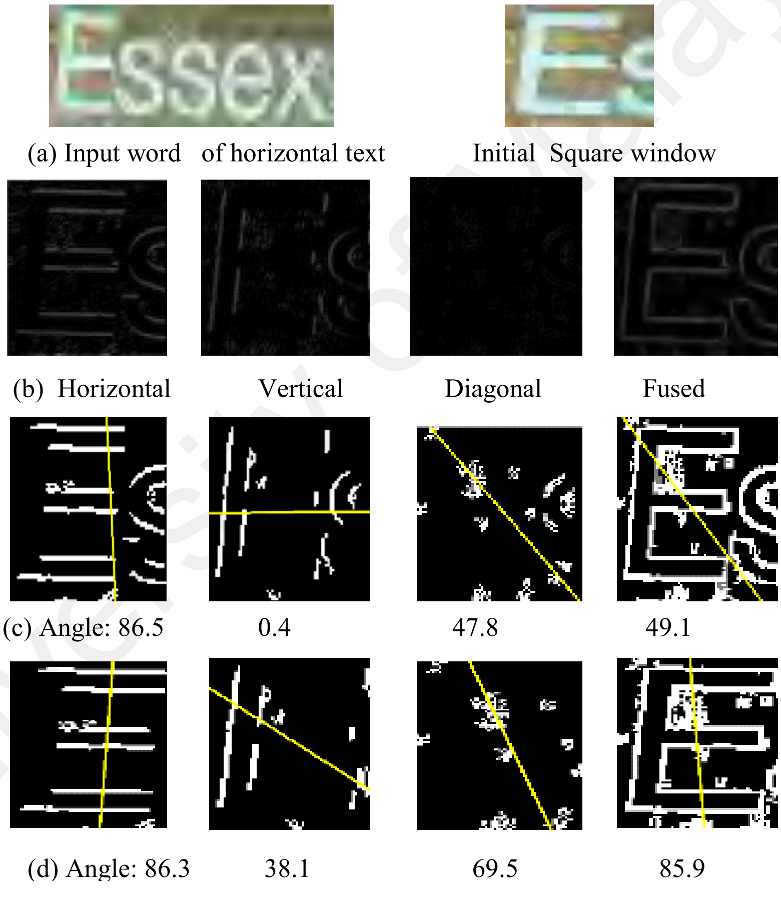
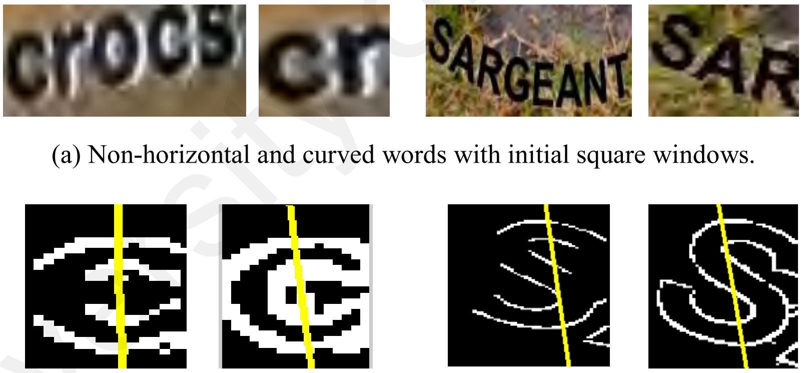


Figure 7.6: Automatic window fixing for horizontal text.

The proposed method iteratively calculates the angle by reducing three pixels at each iteration for both the horizontal window and the fused window until the difference between the horizontal and the fused windows satisfies with + 3 or -3 differences. It can be verified from the example in Figure 7.6 (d), where the last iteration results with angles are shown. The difference between the fused and the horizontal window angles is -0.4, which satisfies the condition with + 3 or -3. These two angles match in the sense that the window contains the exact character without any extra information as shown in Figure 7.6 (d) at fused results. From Figure 7.6 (c) - (d) that the angles of the vertical and the diagonal window do not play a role in calculating angles. In this way, the proposed method determines the correct window iteratively with the help of angle information. The same thing is true for non-horizontal and curved text as portrayed in Figure 7.6 (a) - (b). In some situation, there may have the exact vertical direction. In this case, instead of considering the horizontal frequency sub-band, the vertical sub-band is considered to find matches with the fused window for angle calculation. In addition, the procedure terminates with the angle of zero degrees rather than 90 degrees. The algorithmic steps of the iterative process for finding window are described in Algorithm-7.1.



(a) Non-horizontal and curved words with initial square windows.

(b) 88.7 (H) 85.6 (F) 85.1 (H) 83.1(F)

Figure 7.6: Automatic window fixing for non-horizontal and curved text: (b) is the last result of iterative algorithm for Horizontal (H) and Fused (F) sub-bands.

Algorithm 7.1: Automatic Window Selection

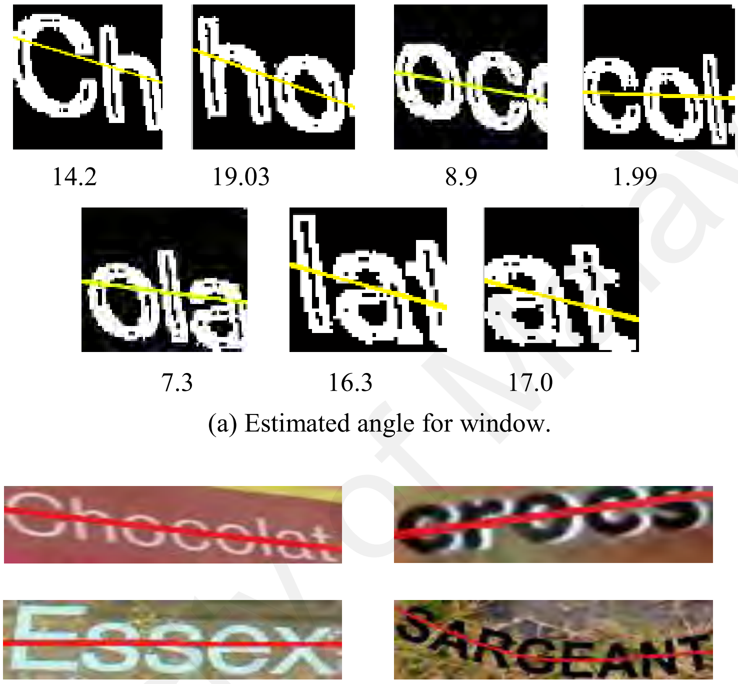
```
Input:   Detected text image of width $w$ and height $h$

1.    For text image, do the following:

      A. Take first Square of window $W_{init}$, where $W_{init} = \{P_1, P_2, \ldots, P_n\}$ is image of width $h$ and height $h$.
      B. Apply wavelet decomposition on $W_{init}$ to generate 3 sub-bands HL, LH, HH and subsequently a fusion image Φ using union operator on 3 sub-bands.
      C. Obtained text cluster ($HH_{txt}$ , $\Phi_{txt}$) and non-text cluster from HH and Φ using K-means.
      D. Compute the covariance matrix ($S$) using $\frac{1}{m}\sum_{i=1}^{n} x^{(i)}(x^{(i)})^T$ of $HH_{txt}$ and $\Phi_{txt}$ where   $x$ is coordinate of text pixel, $m$ is the total number of pixel, $T$ is the transpose matrix.
      E. Apply PCA on  $S$ and find $Z$ direction vector having maximum information.
      F. Estimate PCA angle ($\theta_1$) of HL matrix using $\sinh^{-1}(\frac{z_{12}}{\sqrt[2]{z_{11}^2+z_{12}^2}})$ and PCA angle ($\theta_2$) of Φ matrix using$\sinh^{-1}(\frac{z_{22}}{\sqrt[2]{z_{21}^2+z_{22}^2}})$.
      G. While the absolute difference between  $\theta_1$ and $\theta_2$ is greater than 3,
         a. $W_{init} = W_{init} - \{p_1, p_2, p_3\}$;
         b. Use step (B) ~ step (F) to calculate the $\theta_1$ and $\theta_2$.
      H. Apply shrinking and expand algorithm to generate a sequence of window size W=[$W_1$ $W_2$ ... $W_n$].
      I. For each window
         a. Extract feature vector.
         b. Apply SVM classification on feature vector to get confidence score ($q$).
      J. Select the window $W_{op}$ having the maximum SVM score among all windows using  $W_{i:q=\max\{q_{1,}q_{2,}\ldots q_n\}}$
      K. Recognize the text in $W_{op}$
      L. While the window is not at the end of a text region, do the following:
         a. Move the window with the direction of $\frac{\theta_1+\theta_2}{2}$
         b. Use step (A) ~ step (K) to generate angle and optimal window.

Output: Optimal character window.
```

In Algorithm-7.1, W, W_{init} denote the recognized window pixel matrix, while HL, LH, and HH are the high-frequency sub-bands of wavelet decomposition on W, Φ is the fused band, Z denotes the direction vector determined by PCA, θ_1 denotes the PCA angle of HL matrix, and θ_2 denotes the PCA angle of Φ matrix. Variable X is the feature vector generated by features, and q denotes the SVM classifier score for each feature vector X.



(a) Estimated angle for window.

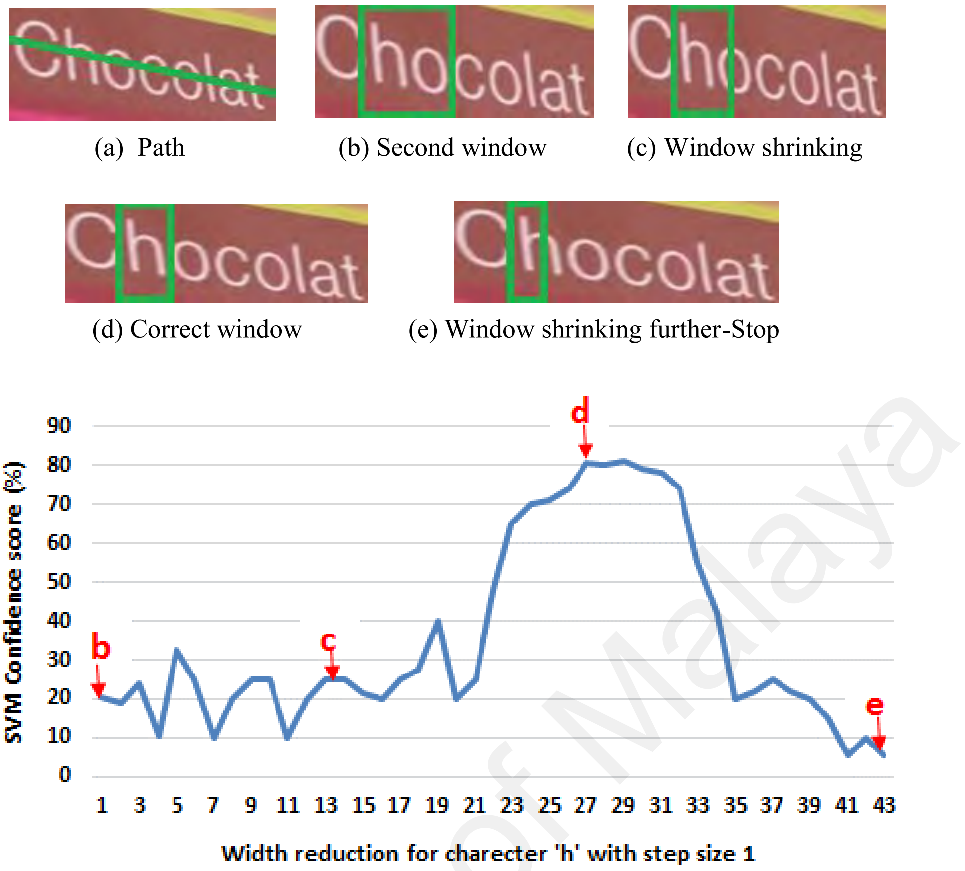
(b) Paths for the different oriented texts.

Figure 7.7: Examples of path estimation for the arbitrary oriented text using fused results with angles.

With the aforementioned step, the proposed algorithm fixes the appropriate window for the initial character. Next, to move to the next character with the same window size, it is needed to find the direction of the text. For this, the angle of the initial window of the fused result is used as the direction to move over the text. The angle of the next window is then used as the direction to move further. This fixing process continues in a non-overlapping fashion until the window reaches the end of the word as displayed in Figure 7.7 (a). Here, the angles of window movements are observed over a text according to text direction, which results in a path for moving the window as depicted in Figure 7.7 (b).

Sometimes, due to upper or lower case and font size variations, the fixed window for the initial character using an automatic window size determination may not fix for neighbor characters during moving. The same procedure can also be adapted for fixing a correct window for the next moved window. But, this procedure is slightly sensitive to very small fonts, while small fonts are common for this work. Therefore, the same square window in a non-overlapping way is moved to the next character and calculate the confidence score using an SVM classifier to fix the correct window for the character by expanding and shirking the window pixel by pixel. The procedure to use an SVM classifier for calculating confidence score is as follows. For the extracted features, a mapping is done between the window that is moving over character images and its label. This is defined as $x \rightarrow y, \quad where\ x \in X$ is a character in window and $y \in Y$ is its class label. Here $x \in R^n$, where *n* is the dimension of features extracted from the window. For the input set X and output set Y, training set will be $(x_1, y_1), \ldots, (x_w, y_w)$. In the testing phase, for an unknown or query window$x_q \in X$, SVM finds appropriate label$y_q \in Y$. In this work, RBF kernel has been used which is the most popular function that has been used in literature. RBF kernel is function k, such that for all $x_{r,} x_s \in X$, $k(x_r, x_s) = \left(\Phi(x_r). \Phi(x_s)\right)$, where Φ is the mapping from *X* to the dot product feature space. More details regarding training and the kernel of the SVM classifier can be found in (Cortes & Vapnik, 1995).

When a text has uniform sized characters, window fixing using the SVM classifier terminates quickly. This is the advantage of the automatic window determination. Window fixing using the SVM classifier serves as a verifier for character window detection. Before recognition, a set of features, which will be discussed later in the same section, is extracted to calculate a confidence score with the SVM classifier. Since the proposed method considers the same window to move to the next character, it saves a large number of computations as it moves character by character but not pixel by pixel in contrast to the existing methods (Mishra et al., 2012b; S. Roy, Roy, et al., 2013a; K. Wang et al., 2011b). When a text has uniform sized characters, window fixing using the SVM classifier terminates quickly. This is the advantage of the automatic window determination. Window fixing using the SVM classifier serves as a verifier for character window detection. More details regarding the training and the kernel of the SVM classifier can be found in (El-Yacoubi, Gilloux, Sabourin, & Suen, 1999).When the confidence score gives the maximum with ground truth, it is considered as the actual window for the character. Figure 7.8 portrays the process, where (a) provides the trajectory, (b) shows the second window, (c) shows the shrinking window, (d) shows the correct window of character "h", (e) shows the shrinking window further, and (f) shows the confidence score given by SVM for window reduction. Figure 7.8 (f) shows the maximum confidence score for fixing the correct window, and the confidence score decreases when the window reduces further.



(a) Path (b) Second window (c) Window shrinking

(d) Correct window (e) Window shrinking further-Stop

Figure 7.8: Character detection using confidence score of SVM.

7.3.2 HMM-based Method for Video Text Recognition

The previous section provides a window to traverse a word character by character through path estimation of any direction. For each window, a new set of features comprising statistical-texture and spatial features is proposed in contourlet wavelet domain. Here, the window is referring one character according to the previous step. Motivated by the alternative review in (Liang et al., 2015a; Shivakumara, Wu, et al., 2017; L. Wu et al., 2015; Yin, Zuo, Tian, & Liu, 2016), a novel method has been proposed to combine the strengths of statistical features which generally help in extracting shapes of characters, texture features which help in extracting character appearance, and spatial features which help in extracting inter and intra symmetrical features of characters components.

For each window (character), the proposed method obtains high-frequency sub-bands. These sub-bands are horizontal, vertical and diagonal, generated with the help of contourlet wavelets (Haar). For each sector of each window of the word image, statistical, textural and run length based features are extracted as defined in (P. Agrawal, Vatsa, & Singh, 2014), where the definitions and formula are presented. These features are used for complex mammogram identification successfully in (P. Agrawal et al., 2014). In this work, for each sector, 90 statistical-textures and 32 run-length based features are extracted. For 8 sectors in three high-frequency sub-bands with two levels of decomposition, totally 122×8×3×2=5856 features are extracted, and then fed to HMM for recognition.

Hidden Markov Model (HMM) is explored as it has the ability to recognize words of different forms (S. Roy, Roy, et al., 2013a). This is because it extracts context information by studying the spatial information of characters. A brief explanation of HMM for recognizing words follows as: A feature vector sequence $X = X_1X_2...X_N$ is first generated from multi-oriented words and is processed using Hidden Markov Models (HMMs). One of the advantages of HMMs is that they are able to cope with variable-length data. The basic models are character models. The ground-truth of text line and feature vectors are jointly trained in supervised fashion to build character model. In order to generate the word models, character models are merged using the ground-truth. Viterbi algorithm generates the best character sequence. This algorithm employs likelihood estimation for recognition. The popular HTK toolkit (Young, Jansen, Odell, Ollason, & Woodland, 1995) has been used for implementing HMMs,. For finding the values of parameters, the same instructions have been followed given in (S. Roy, Roy, et al., 2013a). More details about HMM and parameter setting can be found in (S. Roy, Roy, et al., 2013a).

7.4 Experimental Results and Comparative Study

The recognition results on standard and collected dataset are presented in this section. Section 7.4.1 presents the result of binarization based approach. In section 7.4.3, the result of classifier-based approach is demonstrated. Finally, in Section 7.4.4 comparative study of existing methods against proposed method are analyzed.

7.4.1 Datasets and Evaluation

The proposed Bayesian-based method is evaluated on two types of dataset, namely, word images collected from different video and words from natural scene images have been used. In video data, the whole data has been split into two sub-data as horizontal text data and non-horizontal text data (this includes texts in the curved array). The curved data generally consists of different oriented characters and words while horizontal and non-horizontal straight data consists of unidirectional characters and words. Therefore, handling curved data is hard. Those data are named as horizontal video data and NUS data when reporting the experimental results. Besides, a standard data (Hua's) which is available publicly has been used, and it contains 45 video images having horizontal text lines (Hua, Wenyin, & Zhang, 2004). For scene text data, six datasets, namely, ICDAR 2003 (Lucas et al., 2003), ICDAR 2011 (Shahab, Shafait, & Dengel, 2011), BDIII (Karatzas, Mestre, Mas, Nourbakhsh, & Roy, 2011), (K. Wang & Belongie, 2010), PAMI-2009 (Weinman, Learned-Miller, & Hanson, 2009) and MSRA TD500 (Yao et al., 2012), which are available publicly have been considered. ICDAR 2003, ICDAR 2011 and BDIII datasets are popular; thus considered as benchmark datasets. The majority of text lines in these datasets appeared in the horizontal direction, but images have the complex background like greenery, leaves, etc. The SVT dataset contains mainly street view data, and text lines are in the horizontal direction. The PAMI 2009 dataset contains horizontal as well as non-horizontal text lines without a large variation, and the MSRA TD500 dataset contains multi-oriented text lines with lots of varieties.

In summary, 300 horizontal, 300 multi-oriented video text lines are collected from TRECVID video database, and 45 Hua's data are considered for evaluating the proposed method on video data. For natural scene data, 918 text lines from BDIII, 1110 text lines from ICDAR 2003, 716 text lines from ICDAR 2011, 647 text lines from SVT, 215 text lines from PAMI 2009, 310 text lines from MSRA-TD500 are considered. In total 4561 text line images which include 690 video text lines images, 3916 scene text line images are experimented. The performance of the proposed algorithm is evaluated by word recognition rate which is measured 'before binarization' and 'after binarization' given by the Tesseract OCR for the video data. For scene text data, word recognition rate and pixel level accuracy have been measured. For these measurements, there are some metrics namely, (1) Recall (R), which is ratio of a total number of matched binary pixel divided by total number pixels of the word. It describes how many pixels have been correctly detected out of the total number of pixels. (2) Precision (P), which calculates how many detected pixels are matched out of the total number of detected pixels. (3) F-Measure (F-M), which is defined as the harmonic mean of recall and precision. (4) p-Measure (p-M), which is also harmonic mean of pseudo- Recall and precision. (5) pseudo-Recall (p-R), which is computed based on skeletonized ground truth image with respect to the resultant binary image. A detailed description of these parameters can be found in (D. Kumar, Prasad, & Ramakrishnan, 2012; Pratikakis, Gatos, & Ntirogiannis, 2012). Since the ground truths at the pixel level for all the six mentioned databases except MSRA-TD500 are available in (Yao et al., 2012), the proposed technique is evaluated in terms of pixel level accuracy on the five databases. Note that pixel accuracy measure is not calculated for other video data since their ground truths are not available. Therefore, word recognition rate has been reported for video data.

The Bayesian-based approach is compared with six baseline methods that are: Niblack (Niblack, 1985), Otsu (Otsu, 1979), Wolf (Wolf et al., 2002), WGF (S. Roy et al., 2012), ABM (Chattopadhyay et al., 2013b) and ASB (Moghaddam & Cheriet, 2010). Wolf (Wolf et al., 2002) uses texture features and improved thresholding technique for binarization. An automatic selection of binarization system was proposed in ABM (Chattopadhyay et al., 2013b) where the selection is completely automatic and guided by the machine learning approaches. It suggests a set of one or more binarization techniques suitable for different regions of the input image. ASB (Moghaddam & Cheriet, 2010) makes use of Otsu method for developing adaptive binarization method (AdOtsu) for degraded documents using restoring weak connections and strokes. The reason to choose the first three existing methods (Niblack, 1985; Otsu, 1979; Wolf et al., 2002) is that these are often considered as baselines methods for binarization. The later three existing methods (Chattopadhyay et al., 2013b; Moghaddam & Cheriet, 2010; S. Roy et al., 2012) are recent methods, and they address the problems of contrast and background variations of the images, which is similar to video images.

To evaluate the SVM-HMM based recognition method, standard databases, such as ICDAR 2013 video, ICDAR 2011, SVT scene data, ICDAR 2011 born digital data and South Indian data are considered for experimentation. For all the recognition experiments, the ground truth has been used for calculating recognition accuracy. In addition, the standard measures as presented in (Phan et al., 2013) are followed at both word and character levels for calculating recognition accuracy. This consists of four experiments, namely, evaluating the proposed method on video which comprises ICDAR 2013 data, natural scene which comprises ICDAR 2011, SVT data, Born digital which comprises ICDAR 2011 data, and South Indian data which is created by us. For evaluating the proposed method, respective ground truth and testing data reported in the databases, are used for calculating recognition rates for all the experiments in this work. However, for South Indian Data (SID), 200 words which include 50 for each language is used to validate as in the line of SVT data. HMM is trained according to the guidelines given in

(Young et al., 1995) to obtain the values for the parameters. For training, 750, 1850, 1700, 2200 and 2540 (650 for Kannada, 550 for Malayalam, 700 for Tamil and 640 for Telugu) words are used for ICDAR 2013 (I2013) video, ICDAR 2011 (I2011) scene, SVT, ICDAR 2011 (I2011) Born Digital (BD) and South Indian Data (SID), respectively. In total, the proposed approach considers 9040 words for training in this work.

To show the effectiveness of the classifier-based text recognition method, the state of the art methods have been implemented for text recognition in scanned images, natural scene images, and videos. For example, Milyae et al. (Milyaev et al., 2013) and Howe (Howe, 2013) recognize text through binarization in natural images, Su et al. (Su et al., 2013) recognize text through binarization in degraded document images, Roy et al. (S. Roy, Roy, et al., 2013a) recognize text without binarization in natural scene image, Roy et al. (S. Roy et al., 2012) recognize text through binarization in video, and Jaderberg et al. (Jaderberg et al., 2016) explore deep learning in natural scene images. Lee and Kim (S. J. Lee & Kim, 2016) has also proposed deep convolutional neural network for slab number recognition. Note that the codes or exe files are available for Milyae et al., Howe et al., Jaderberg et al., Lee and Kim and Su et al.'s methods, while the other two approaches proposed by Roy et al. are implemented according to their papers. These methods are chosen because it involves binarization, classifiers and deep learning for recognition as the proposed method to give fair comparative studies.

7.4.2 Experiments on Binarization based Approach

Experiment is conducted on the same 100 samples chosen randomly from different databases in order to test the method's performance without using probability. For this purpose, k-means clustering algorithm has been applied to the integrated enhanced image to classify text cluster. Then normalization has been done on the values in the text cluster to the range between 0 and 1. The method fixes 0.5 as a threshold for the pixels in text cluster to classify non-text and text pixels. To obtain recognition rate, the results are sent to Tesseract OCR. Word recognition rate, F-M, p-M, R, P, and p-R are calculated. The results are reported in Table 7.1. From the table, it is noted that proposed algorithm without probability yields very poor accuracy in terms of all measures compared to the same with probability. Therefore, it is evidence that the use of probability makes a significant difference in obtaining good accuracy.

Similarly, in order to analyze the combinations of fusion criterion with the different domains in different ways, experiments are conducted on the same samples to calculate the performance measures, and the results are reported in Table 7.1. The combinations of wavelet with gradient, gradient with RGB, and RGB with wavelet have been tested to realize the contribution of each combination. Table 7.1 shows that the combinations except for RGB with Wavelet give poor results than the results of the proposed method which integrates RGB, wavelet, and gradient. It is also observed from Table 7.1 that RGB with Wavelet gives better precision than the proposed method. However, the same combination gives worst recall and other measures. Therefore, it can be asserted that the use of three domains and for binarization by integrating them is better than any other combinations.

Table 7.1: Pixel level accuracy of combinations of fusion criteria (in %).

Methods	Word	F-M	p-M	R	P	p-R
Without probability	12.1	30.0	32.2	24.5	38.6	27.6
Wavelet + Gradient	14.4	25.7	28.8	20.1	35.7	24.2
Gradient + RGB	8.3	29.1	30.9	23.4	38.6	25.8
RGB + Wavelet	9.2	20.6	19.7	12.7	54.5	12.0
Proposed Method	39.1	55.6	51.5	59.4	52.3	50.7

7.4.3 Experiments on Classifier-based Approach

The system with 2.59 GHz, 8GB RAM and Window 8 is used for experimentation. Average processing time, computed from the mean of processing 100 data, has been reported in Table 7.2. These data are chosen randomly from the databases. Table 7.2 shows that the feature extraction with HMM for recognition consumes more time because HMM process requires more computations. Table 7.2 shows that the proposed method takes, on an average, more processing time for video data and MSRA data compared to the other databases. This is valid because video involves processing of temporal frames and MSRA involves arbitrarily-oriented text which requires more computations compared to horizontal text. Overall, the proposed method consumes a few seconds for each image in order to recognize text in the image. This is due to MATLAB software. It is also noted that the processing time depends on many factors, such as data structure of the algorithm, system configuration, and platform. Since the main aim of the proposed work is to develop a generic method for recognizing text irrespective orientation, contrast variations, scripts, etc., prototype or working model development is considered as beyond the scope of this work.

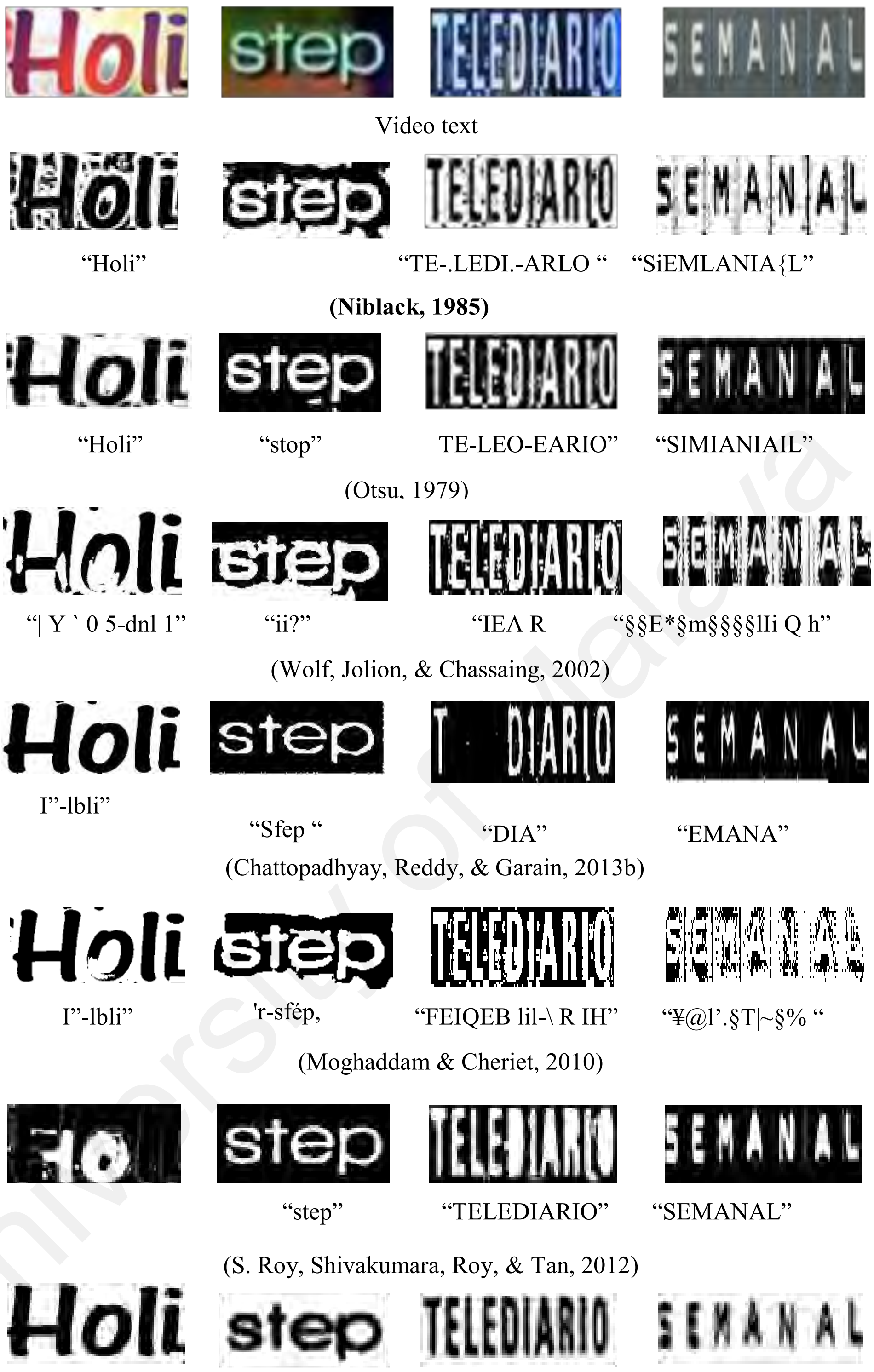
Table 7.2: Average Processing Time of the proposed method for recognition on different databases in seconds.

Databases	Recognition		Total
	AWD	Features+HMM	
ICDAR 2015 Video	3.2	3.9	7.1
YVT Video	3.1	3.5	6.6
ICDAR 2013 Scene Data	2.7	3.4	6.1
SVT Scene Data	2.9	3.0	5.9
MSRA Scene Data	3.3	3.9	7.2
ICDAR 2011 Born Digital Data	2.6	3.2	5.8
South Indian Data	3	1.9	4.9

7.4.4 Comparative Study and Discussion

Sample qualitative results of the proposed and existing methods are shown in Figure 7.9 for collected video and Hua's data. In Figure 7.9, the input images are shown in the first row, the binarization results of six existing methods and the proposed method are shown, respectively from the 2nd row to the 8th row. The OCR results are displayed under every image within double quote for all the methods. It is noticed from the results in Figure 7.9 that Niblack, Wolf, Otsu, ABM, ASB recognize the first image of collected data properly and fail to recognize the second-word image because of the complex background of the image. WGF recognizes the second word but fails for the first word. Moreover, the proposed technique recognizes both the words correctly. The same inference can be drawn from the results shown for Hua's data in Figure 7.9 except for the proposed method and WGF. All the state-of-the-art fail to recognize the input words properly. Sample qualitative result of the proposed and existing method on curved data is shown in Figure 7.10. Figure 7.10 shows that all existing methods give much poorer results than the proposed method for the curved input images due to their inability to handle curved data. It is noted from Figure 7.10 that since the OCR engine accepts only horizontal characters, it gives nothing for the multi-oriented data. Therefore, the proposed algorithm is better than existing approaches irrespective of orientation and text type.

The quantitative results are reported for both horizontal (this includes Hua's data also) and curved video data in Table 7.3 where Tesseract OCR has been applied on input images directly, called 'before binarization' (BB) and on the output of binarization methods, called 'after binarization' (AB) to calculate character recognition rate. Table 7.3 shows that character recognition accuracies before binarization is lower than after binarization for all the methods including the proposed method. This shows that applying OCR direct on images may not be useful. When the recognition results of previous method (wavelet-gradient-fusion WGF) with the existing methods have been considered, the previous method's performance is better than other state-of-the-arts. The proposed algorithm gives superior result than all the existing techniques including the WGF. It can also be inferred that from Table 7.3 that all methods provide good accuracy for horizontal data but generate poor accuracy for curved data because of orientation problem. In addition, OCR recognizes horizontal character well but not multi-oriented images to recognize the characters.



Video text

“Holi” “TE-.LEDI.-ARLO “ “SiEMLANIA{L”

(Niblack, 1985)

“Holi” “stop” TE-LEO-EARIO” “SIMIANIAIL”

(Otsu, 1979)

“| Y ` 0 5-dnl 1” “ii?” “IEA R “§§E*§m§§§§lIi Q h”

(Wolf, Jolion, & Chassaing, 2002)

I”-lbli” “Sfep “ “DIA” “EMANA”

(Chattopadhyay, Reddy, & Garain, 2013b)

I”-lbli” 'r-sfép, “FEIQEB lil-\ R IH” “¥@l’.§T|~§% “

(Moghaddam & Cheriet, 2010)

“step” “TELEDIARIO” “SEMANAL”

(S. Roy, Shivakumara, Roy, & Tan, 2012)

“Holi” “step” “TELEDIARO” “SEMANAL”

Proposed Method

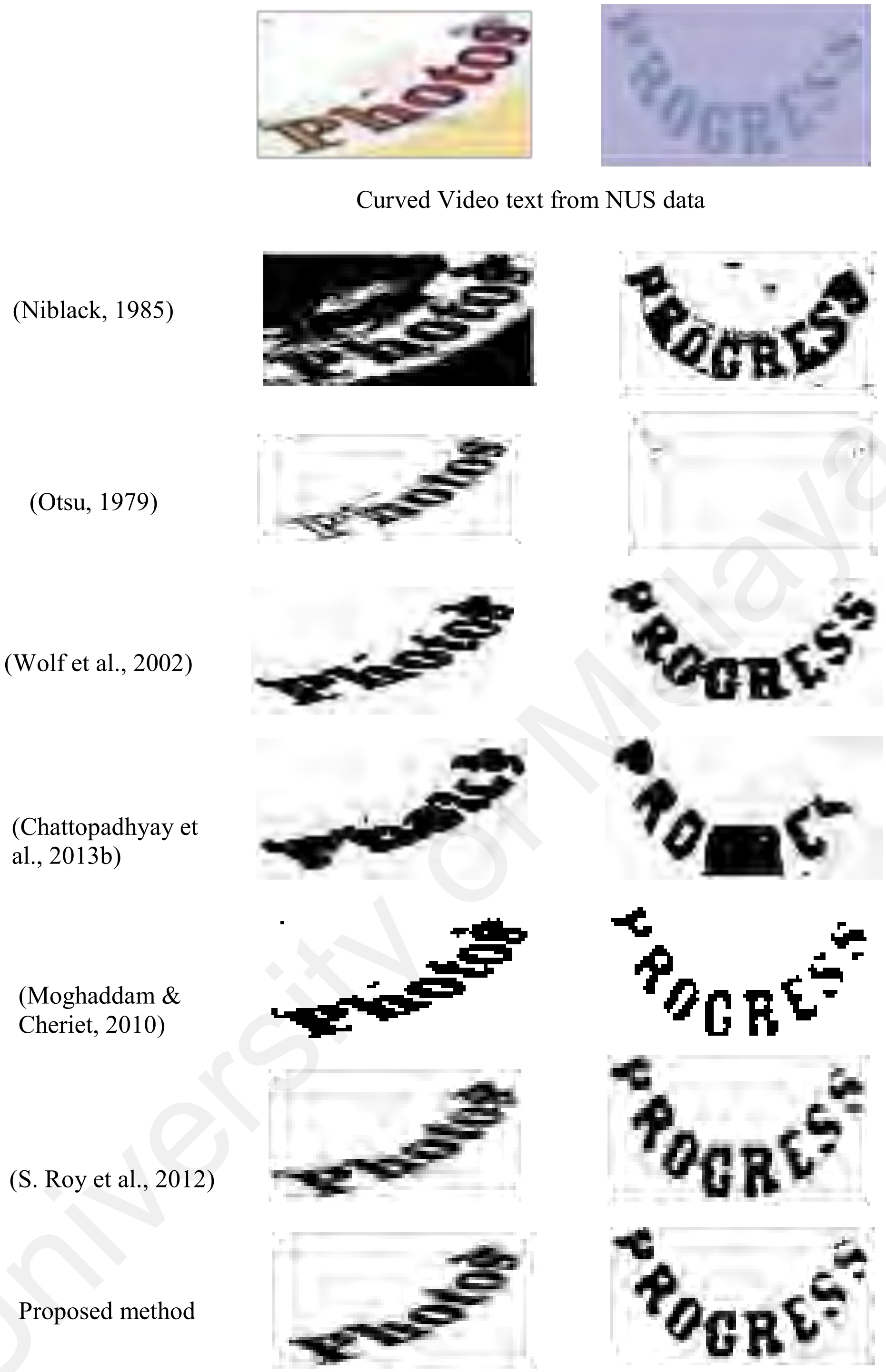
Figure 7.9: Binarization and recognition results of the proposed and existing methods on collected data and Hua’s data.

Table 7.3: Recognition rate of the proposed and existing methods before and after binarization with Tesseract OCR on video dataset (in %).

(BB denotes 'before binarization' and AB denotes 'after binarization')

Methosds	Collected (horizontal)		NUS (Non- horizontal)	
	BB	AB	BB	AB
(Niblack, 1985)	36.12	44.18	6.45	11.63
(Otsu, 1979)		41.18		15.56
(S. Roy et al., 2012)		50.18		19.71
(Chattopadhyay et al., 2013b)		42.13		17.32
(Wolf et al., 2002)		46.23		18.37
(Moghaddam & Cheriet, 2010)		41.37		17.82
Proposed Method		56.18		21.12

Sample qualitative results for the six datasets of natural scene text are shown in Figure 7.10 and Figure 7.11 where input images of different orientation are shown in the first row, binarization results and recognition results of the six existing methods and the proposed method are shown from 2nd row to 8th row. From Figure 7.11, it is observed that existing methods fail to recognize the input images properly due to complex background scene images of ICDAR 2003 and 2011 data while the proposed algorithm yields good results. It is noticed from Figure 7.10 that the result generated from Otsu technique is better than other existing works but worse than the proposed one. Figure 7.11 shows that Otsu technique works well on PAMI 09, BDI11 and SVT data but it is worse than WGF and the proposed method. Figure 7.11 also shows that all existing methods including WGF and Otsu cannot handle multi-oriented text (MSRA) because of the non-horizontal nature while the proposed method gives good results. This shows that the proposed method is good for different data and works well irrespective of orientation and background variations.



Curved Video text from NUS data

Figure 7.10: Binarization and recognition results of the proposed and existing methods on curved data of NUS.

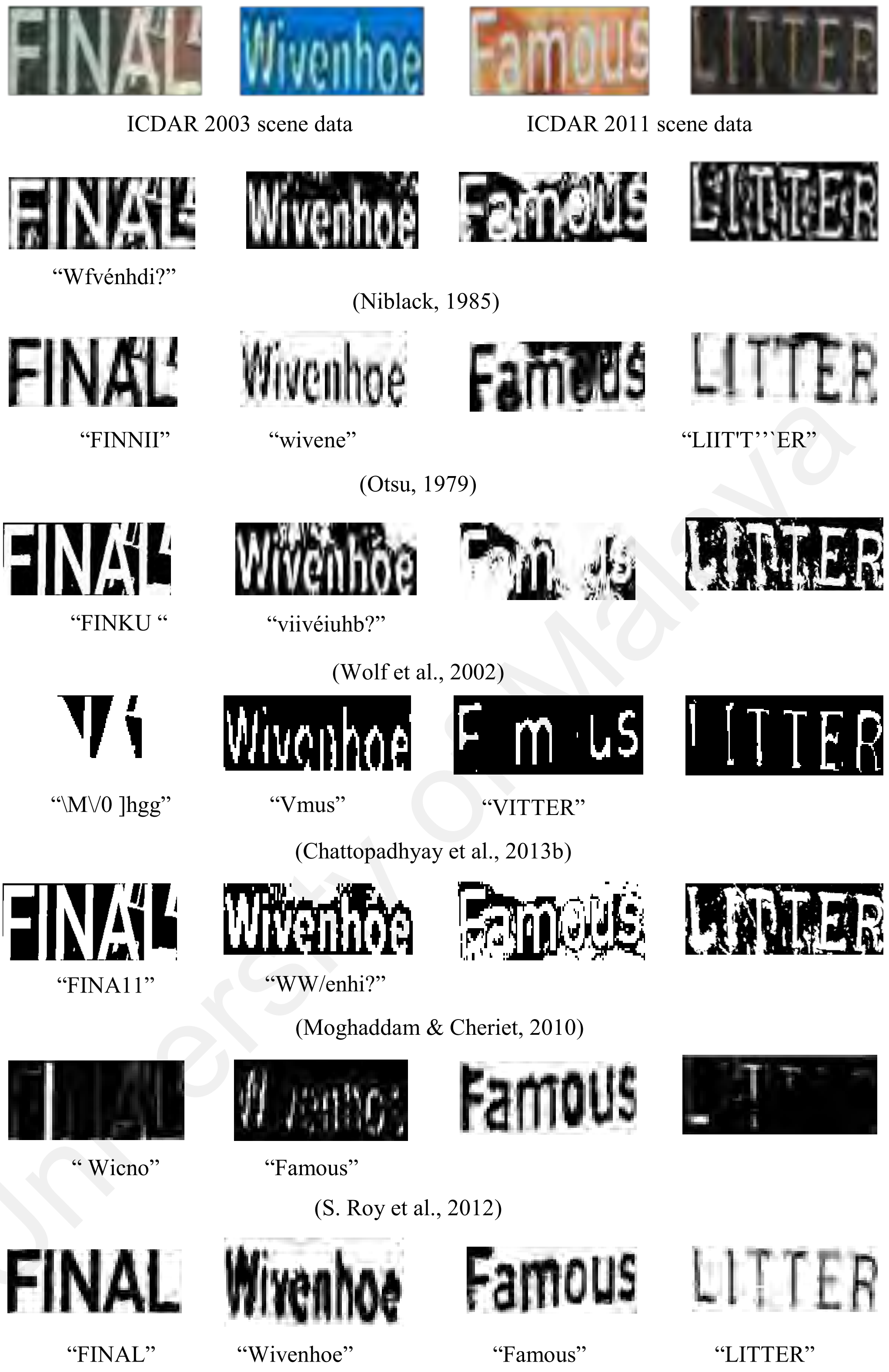
The quantitative results of the proposed and existing methods are reported for the six datasets of natural scene text in Table 7.4 where the word recognition rate before binarization (BB) given within brackets in the first row and after binarization (AB) are reported. The recognition accuracy before binarization is lower than after binarization especially for SVT and curved dataset. All the methods report low recognition rate due to the complexity of the background and orientation. The previous method is better when comparison is done between previous method (WGF) and the state-of-the-art methods. Furthermore, the proposed algorithm performs best than rest including WGF. This is because of the Bayesian classifier and the way a priori probability is estimated with different ways of enhancement methods. The major drawback of the previous works is that these methods generally use thresholds for binarization and those thresholds are derived based on plane background images rather than considering the background of complex natural scene as well as video.

Table 7.4: Recognition rate of the proposed and existing methods before and after binarization with Tesseract OCR on scene dataset (in %).

(I-03 denotes ICDAR 2003 and I-11 denotes ICDAR 2011 dataset)

Methods	BD11 (34.5)	I-03 (32.1)	I-11 (27.3)	PAMI (51.1)	SVT (16.3)	MSR (27.1)
	A.B	A.B	A.B	A.B	A.B	ABF
(Niblack, 1985)	53.1	57.1	54.6	33.1	53.1	25.2
(Otsu, 1979)	53.1	56.1	52.1	53.2	53.3	31.1
(S. Roy et al., 2012)	54.5	57.6	58.2	53.1	51.5	33.3
(Chattopadhyay et al., 2013b)	34.2	46.0	35.5	38.0	32.2	27.2
(Wolf et al., 2002)	40.2	42.6	34.3	48.8	28.3	32.2
(Moghaddam & Cheriet, 2010)	38.3	40..5	35.0	48.3	30.3	31.3
Proposed Method	57.1	61.3	58.6	57.1	52.1	38.2

The pixel level accuracy generated by evaluation software is reported to evaluate the binarization results as nowadays these measures are quite popular for scene text data.



(Niblack, 1985)

(Otsu, 1979)

(Wolf et al., 2002)

(Chattopadhyay et al., 2013b)

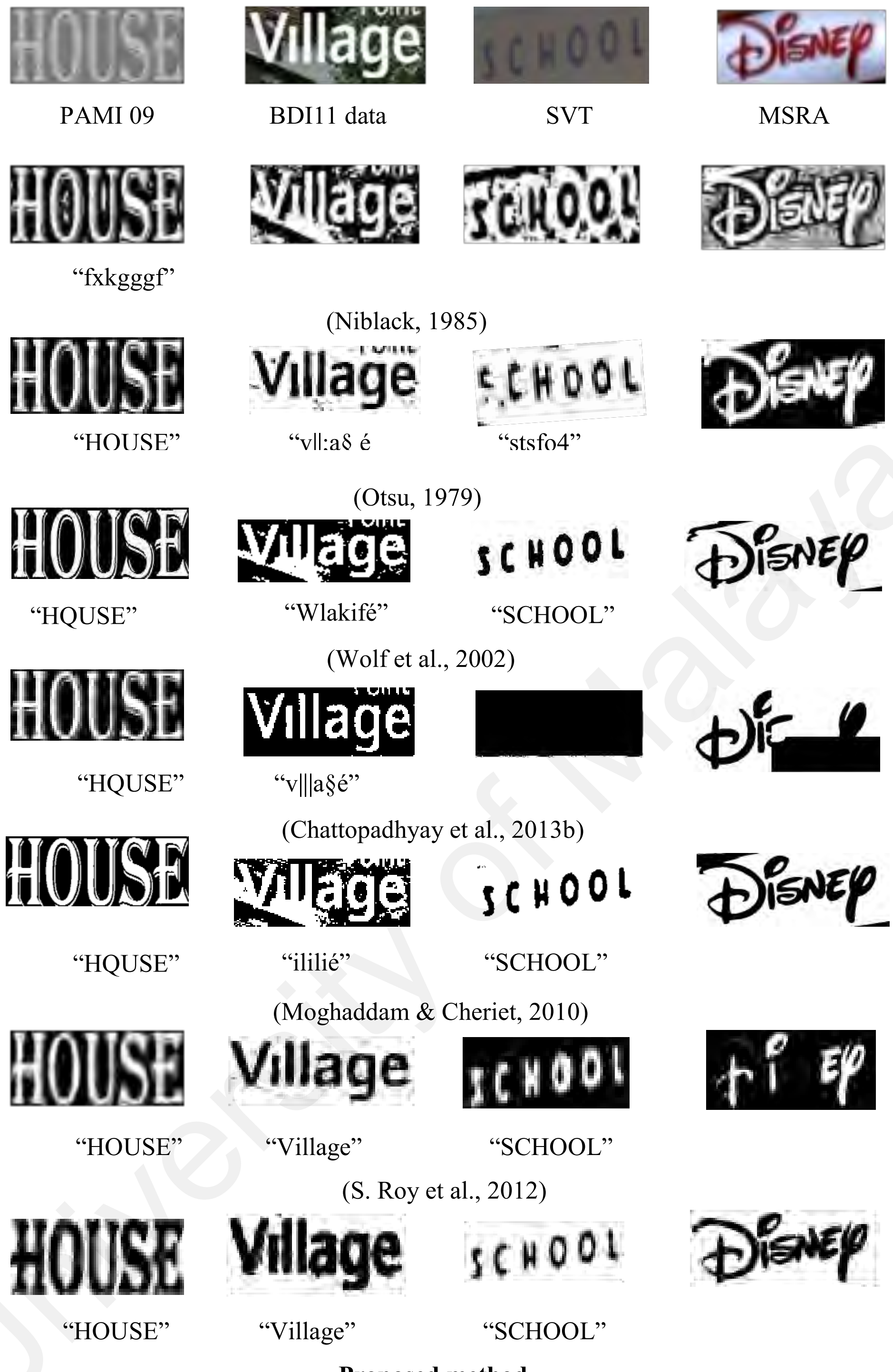
(Moghaddam & Cheriet, 2010)

(S. Roy et al., 2012)

Proposed method

Figure 7.11: Binarization and recognition results of the proposed and existing methods for ICDAR 2003 and ICDAR 2011 scene data.

Table 7.5 shows the results of all the existing methods on different parameters for all datasets. When the comparison is done of the results in Table 7.5, the proposed algorithm generates superior results from other recent methods in terms of pixel level accuracy. It can also be observed that for SVT data the method including the proposed method gives poor results because SVT dataset contains complex background and many variations in backgrounds. Therefore, the proposed method is good in case of video as well as natural scene data in terms recognition rate and pixel level accuracy.



(Niblack, 1985)

(Otsu, 1979)

(Wolf et al., 2002)

(Chattopadhyay et al., 2013b)

(Moghaddam & Cheriet, 2010)

(S. Roy et al., 2012)

Proposed method

Figure 7.12: Binarization and recognition results of the proposed and existing methods for PAMI, BDIII, SVT and MSRA scene data.

Table 7.5: Pixel level accuracy of the proposed and existing methods (in %).

(I-03 denotes ICDAR 2003 and I-11 denotes ICDAR 2011 dataset)

Measures	Methods	BDI11	I-03	I-11	PAMI	SVT
F-M	(Niblack, 1985)	39.5	43.0	44.6	50.1	47.6
	Moghaddam & Cheriet, 2010)	55.8	56.0	53.0	50.3	43.7
	(Otsu, 1979)	42.1	43.6	39.8	51.0	36.5
	(Chattopadhyay et al., 2013b)	31.9	40.3	30.2	39.7	38.8
	(Wolf et al., 2002)	57.5	49.5	43.9	52.3	56.4
	(S. Roy et al., 2012)	69.2	52.9	52.0	56.7	51.1
	Proposed Method	82.7	59.3	57.7	61.2	58.2
p-M	(Niblack, 1985)	37.9	41.0	43.4	47.2	45.5
	Moghaddam & Cheriet, 2010)	59.2	56.6	54.6	50.4	44.5
	(Otsu, 1979)	40.2	43.0	38.8	46.5	34.5
	(Chattopadhyay et al., 2013b)	33.4	42.3	25.2	38.3	37.7
	(Wolf et al., 2002)	63.3	50.5	43.5	47.9	58.9
	(S. Roy et al., 2012)	68.1	50.7	51.2	55.6	50.3
	Proposed Method	83.3	59.5	57.1	51.8	58.7
R	(Niblack, 1985)	29.8	34.4	36.0	46.5	35.9
	Moghaddam & Cheriet, 2010)	60.1	51.4	40.8	45.7	33.6
	(Otsu, 1979)	33.2	37.1	34.6	50.5	27.6
	(Chattopadhyay et al., 2013b)	21.3	37.8	19.8	32.0	31.1
	(Wolf et al., 2002)	47.2	39.2	36.6	51.0	59.4
	(S. Roy et al., 2012)	62.1	45.2	41.3	52.1	42.8
	Proposed Method	75.5	49.8	48.8	57.6	48.2
P	(Niblack, 1985)	65.7	61.0	61.1	56.0	74.9
	Moghaddam & Cheriet, 2010)	50.7	65.1	75.7	62.0	63.3
	(Otsu, 1979)	62.8	55.8	49.3	52.3	57.5
	(Chattopadhyay et al., 2013b)	63.7	43.2	63.1	51.0	51.7
	(Wolf et al., 2002)	73.5	67.0	58.3	51.9	53.0
	(S. Roy et al., 2012)	78.3	63.6	70.4	62.3	63.4
	Proposed Method	92.2	78.2	77.0	75.6	81.6
p-R	(Niblack, 1985)	24.7	33.2	34.9	41.9	35.1
	Moghaddam & Cheriet, 2010)	71.2	52.6	42.7	46.1	3.7
	(Otsu, 1979)	31.4	36.7	33.6	42.3	26.3
	(Chattopadhyay et al., 2013b)	22.6	41.4	15.7	30.7	29.6
	(Wolf et al., 2002)	55.6	40.6	36.2	45.0	63.9
	(S. Roy et al., 2012)	60.3	42.1	40.2	50.2	41.7
	Proposed Method	76.1	50.1	48.2	40.9	47.8

In Table 7.6 and Table 7.7, the performance results for video, scene, born digital and South Indian dataset are tabulated. Table 7.6 and Table 7.7 show that the recognition rates for words are lower than those of characters. This is because the proposed HMM does not involve any post-processing, dictionary, or language models for recognizing words. As a result, if one character misses or fails to recognize correctly, the whole word is considered as wrong results for calculating recognition rates. While this is not the case for character recognition rate where each character contributes to recognition rate calculation. The existing methods which use binarization (Howe, 2013; Milyaev et al., 2013; S. Roy et al., 2012; Su et al., 2013) for recognition score low results compared to the methods which do not use binarization (Jaderberg et al., 2016; C.-Y. Lee, Bhardwaj, Di, Jagadeesh, & Piramuthu, 2014; S. J. Lee & Kim, 2016; Phan et al., 2013; S. Roy, Roy, et al., 2013a; B. Shi, Wang, Lyu, Yao, & Bai, 2016) as reported in Table 7.6. This is because though binarization methods are developed for natural scene text recognition, they fail to preserve character shapes for complex background images. At the same time, the OCR (Smith, 2007)which is available publicly is not robust to variant font style. On the other hand, the methods (Jaderberg et al., 2016; C.-Y. Lee et al., 2014; S. J. Lee & Kim, 2016; Phan et al., 2013; S. Roy, Roy, et al., 2013a; B. Shi et al., 2016) give better results at both word and character levels because they extract their own features and classifiers along with dictionaries, language models for achieving better results. In case of Jaderberg et al. (Jaderberg et al., 2016), the method has been executed without lexicons and synthetic images to calculate measures as the proposed method for fair comparative study. Compared to existing methods, the proposed method gives better performance in recognizing character. This is mainly because of the advantage of determining window size according to character size and moving widow over characters.

Note that Roy et al. (S. Roy, Roy, et al., 2013a) use HMM for recognizing text in natural images, Jadeberg et al. (Jaderberg et al., 2016) and Lee & Kim (S. J. Lee & Kim, 2016) use deep learning which has the ability to handle complex issues. Therefore, the performance comparison is done between the proposed method and above existing methods to show the superiority of multi-lingual ability. The accuracy given in Table 7.7 states that the recognition accuracies of the proposed algorithm are better than those state-of-the-arts at both word and character levels. Determining optimal parameter values for the complex script is hard from the above three existing methods, while the proposed method can be extended to any language as it does not involve any specific parameters and lexicons. Interestingly, the result of South Indian data including all the data (Kannada+Tamil+Telugu+Malayalam) together is lower than the other data. Hence, it is an open issue for the researchers. Overall, the proposed method yields the superior result for detecting and recognizing text without many constraints and is invariant to language, orientation, multi-fonts, multi-size, and multi-type text.

Table 7.6: Recognition rates of the proposed and existing approaches on different datasets at word and character levels (in %). W and C indicate word and character recognition rates, respectively.

Approaches	I2013 Video		I2011 Scene		SVT Scene		I2011 BD	
	W	C	W	C	W	C	W	C
(Milyaev et al., 2013)	51.6	61.6	59.6	68.2	51.7	62.3	54.4	61.4
(Howe, 2013)	20.4	30.4	33.5	39.5	30.8	34.8	34.5	38.5
(Su et al., 2013)	31.4	45.4	34.1	47.3	39.4	39.4	37.4	45.4
(El-Yacoubi et al., 1999)	54.2	65.2	56.0	68.7	52.5	58.5	52.5	62.5
(S. Roy, Roy, et al., 2013a)	53.4	66.4	59.2	59.2	53.8	53.8	55.6	60.6
(Phan et al., 2013)		-		-		67.0		-
(C.-Y. Lee et al., 2014)		-		77.0		80.0		-

Table 7.6: Continued

Approaches	I2013 Video		I2011 Scene		SVT Scene		I2011 BD	
	W	C	W	C	W	C	W	C
(B. Shi et al., 2016)			83.2	-	73.6	-	-	-
(Jaderberg et al., 2016)	62.8	67.2	78.4	84.8	72.3	78.43	60.5	70.4
(S. J. Lee & Kim, 2016)	57.5	62.3	71.3	75.5	64.2	72.4	56.9	65.7
Proposed	65.4	70.8	85.3	86.3	76.3	81.7	67.3	72.3

Table 7.7: Recognition rates of the proposed and existing approaches on South Indian datasets at word and character levels (in %). W and C indicate word and character recognition rates, respectively.

Approaches	Kannada		Tamil		Telugu		Malayalam		All	
	W	C	W	C	W	C	W	C	W	C
(S. Roy, Roy, et al., 2013a)	48.2	53.0	52.7	54.6	47.0	53.4	51.0	52.8	10.0	13.7
(Jaderberg et al., 2016)	51.8	56.3	52.5	56.3	51.2	53.7	52.4	56.2	10.6	14.2
(S. J. Lee & Kim, 2016)	43.2	50.7	46.9	50.2	41.9	45.8	51.5	53.9	8.8	10.6
Proposed	52.3	58.8	53.2	56.8	56.8	59.3	56.4	58.8	15.0	17.5

7.5 Summary

In summary, in this chapter, two video text recognition approaches have been presented. The first approach investigates a Bayesian technique for binarizing variant oriented text in the video. The Bayesian method introduces a new enhancement technique which integrates color, wavelet and gradient bands to obtain an enhanced image. Conditional probabilities and a priori probabilities are estimated without having knowledge of the dataset on the basis of enhancement images. In the second approach, a new idea of determining an automatic window has been introduced according to character size based on the angular relationship between fused and high-frequency wavelet sub-bands. The combination of statistical-textures and spatial information based features are extracted in contourlet wavelet domain for recognition with the help of HMM model. Experimental results depict that the aforementioned methods are able to handle multi-type and multi orientated video and scene text word.

CHAPTER 8: CONCLUSIONS

8.1 Summary of the Proposed Work

In this thesis, the multi-text type recognition in multi-oriented environment consists of video frames categorization, video enhancement, and text detection/keyword spotting and text recognition are discussed.

In chapter 3, the method based on fuzzy-mass and fuzzy-rough are presented for video frames classification of different text types. The combination classifies edge components in each input frame into different groups to extract local information. For each group, the proposed method extracts local and global features, which helps to encode unique relationship for each video class type. Temporal information is used to increase the discriminative power of feature extraction. The extracted features are then fed to a neural network classifier for the final classification.

In Chapter 4, a new enhancement model has been proposed based on Fractional Poisson for increasing fine details in natural scene images as well as video by supprresing the noises introduced by Laplacian operation.

In Chapter 5, the multimodal technique explores the combination of face and skin features for identifying text candidate regions from an input image. For spotting in the video, texture-spatial-features and Cesaro means have been exploited in two different ways. Radon and Fourier transform have been investigated for the text candidates detected by Ceasro means features in gradient domain to extract features locally and globally. Based on features, the proposed approach extracts context information with the help of foreground and background that represents text candidate to eliminate false text candidate, which results in text representatives. Finally, a new minimum cost path estimation has been proposed based on ring growing for restoring the missing information during detection of text representatives, moving along text direction to extract words. For the detected words, the proposed approach extracts the above mentioned local and global features for both foreground and background to perform keyword spotting.

In Chapter 6, tempered based and wavelet-based methods have been proposed for video text type classification respectively, at the line and word level. For line level, a new idea of exploring DCT has been presented for identifying tampering information for the classification of caption and scene texts. For word level, a novel method has been explored based on wavelet and temporal information for the classification of caption and scene texts in the video. The proposed method introduces a new idea of exploring positive and negative coefficients of wavelet decomposition for detecting text candidates. The distribution of text and non-text candidates over caption and scene word images are studied in a novel way to derive four features that give cues for the classification of caption and scene words. Temporal coherency is explored for determining the number of temporal frames to be used and to find stable features to classify caption and scene words correctly.

Finally, video text recognition approaches, one is with binarization, and another is without binarization have been proposed. For binarization based approach a Bayesian classifier-based method has been explored for binarizing multi-oriented text in the video. For without binarization based approach, a new idea has been proposed of determining an automatic window according to character size based on the angular relationship between fused and high-frequency wavelet sub-bands. The combination of statistical-textures and spatial information based features are exploited in contourlet wavelet domain for recognition with the help of HMM model.

8.2 Future Work

The work presented in this thesis is able to recognize the multi-type and multi-oriented text in the video. However, it is observed from experiments results that there are some limitations in aforementioned proposed approaches.

For example, for video classification work, it is noticed from experimental results that the text detection and recognition results decrease when misclassification occurs. Therefore, investigating and introducing an unsupervised method to determine the number of classes will be one objective in the near future. Besides, the presented idea would be extended to more number of classes.

In a similar way, the text detection and recognition accuracies reported in video text enhancement work are still lower compared to scanned documents of the document analysis field. Therefore, future work would be focusing on further improving the enhancement model such that the current text detection and recognition methods can get similar accuracies as in the document analysis field by exploring temporal information in case of video.

In case of text spotting, according to results shown in Figure 8.1, there are still issues with the query words affected severely by blur, fancy fonts, and too low resolution. Therefore, there is a need for a robust method for spotting in complex images. It is seen from the word classification results shown in Figure 8.2 that the proposed method still misclassifies some of the texts that have different fonts and background. Therefore, the future objective will be to extend present text type classification study to address this issue.

The proposed recognition method performs well in English script, but the reported character recognition accuracy is still low for multi-lingual data, too small font, blur, poor quality images, and severely distorted texts, affected by perspective distortion and camera movements as shown in Figure 8.3. The main reason is that the method loses character structure during fixing automatic window for each character and feature extraction. In addition, the recognition method assumes text appears without blur and to some extent clear structure and does not consider occluded text because the method is not capable of restoring missing text information. Therefore, in future for finding a solution to the above-mentioned limitations of the proposed methods, more work will be on exploring robust features based on characteristics of text components which can withstand for recognizing different script in arbitrary orientation in the video.

The investigation will be carried on exploring a deblurring model for the images affected by non-uniform blur. In addition, work planning will be done to extend the proposed recognition method for learning to improve the recognition rate even for different oriented characters and script with the help of the temporal information in the video. The combination of the traditional classifier and deep learning can be proposed for addressing the above issues.

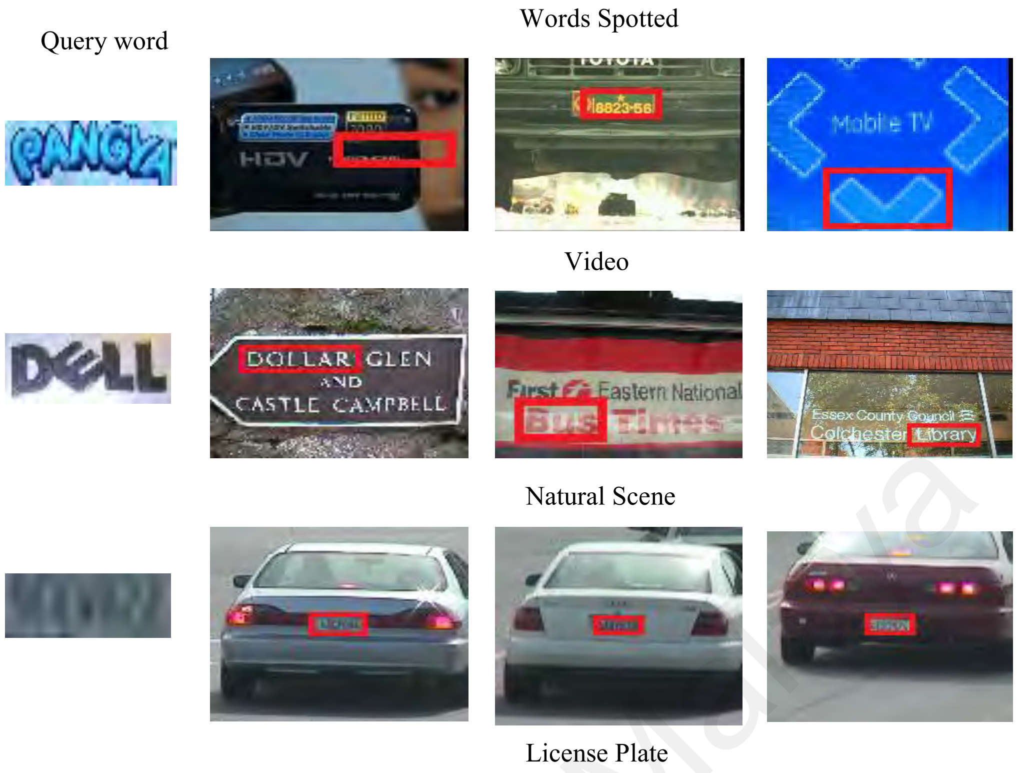
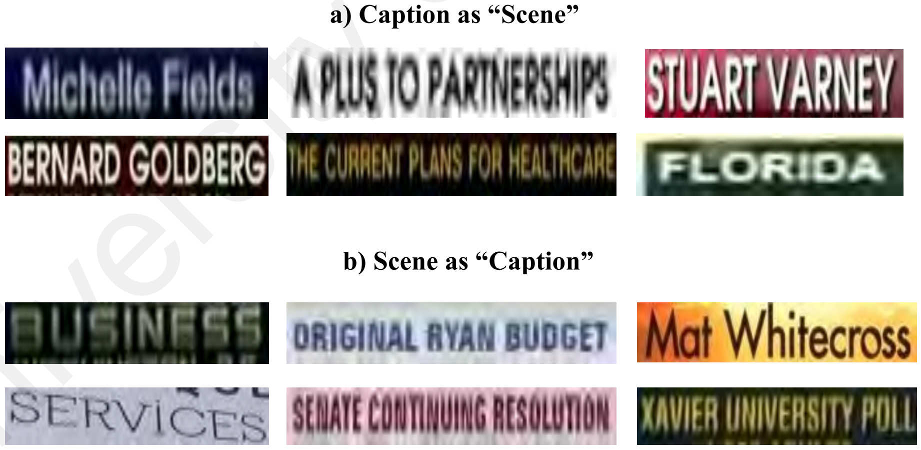

Query word
Words Spotted
Video
Natural Scene
License Plate
DOLLAR GLEN AND CASTLE CAMPBELL
First Eastern National Bus Times
Essex County Council Colchester Library
Mobile TV
HDV

Figure 8.1: Limitations of the proposed approach


a) Caption as “Scene”

b) Scene as “Caption”

Figure 8.2: Samples of unsuccessful classification results of the proposed method. (a) Scene text line misclassified as Caption text line and (b) Caption text line misclassified as Scene text line.

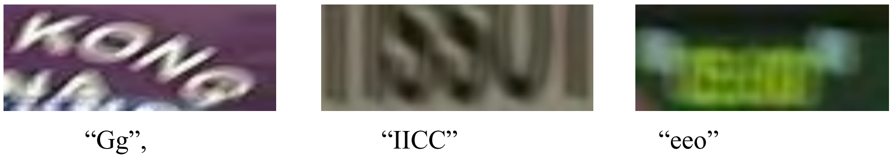


Figure 8.3: Limitation of the proposed text recognition methods.

REFERENCES

Afzal, M. Z., Pastor-Pellicer, J., Shafait, F., Breuel, T. M., Dengel, A., & Liwicki, M. (2015). *Document image binarization using LSTM: a sequence learning approach.* Paper presented at the Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing.

Agrawal, A., Mukherjee, P., Srivastava, S., & Lall, B. (2017). Enhanced Characterness for Text Detection in the Wild. *arXiv preprint arXiv:1712.04927*.

Agrawal, P., Vatsa, M., & Singh, R. (2014). Saliency based mass detection from screening mammograms. *Signal Processing, 99*, 29-47.

Ahmad, R., Rashid, S. F., Afzal, M. Z., Liwicki, M., Dengel, A., & Breuel, T. (2016). *A novel skew detection and correction approach for scanned documents.* Paper presented at the DAS 2016, 12th Intl IAPR Workshop on Document Analysis Systems, At Santorini, Greece.

Ahmadi, E., Azimifar, Z., Shams, M., Famouri, M., & Shafiee, M. J. (2015). Document image binarization using a discriminative structural classifier. *Pattern Recognition Letters, 63*, 36-42.

Almazán, J., Gordo, A., Fornés, A., & Valveny, E. (2014a). Segmentation-free word spotting with exemplar SVMs. *Pattern Recognition, 47*(12), 3967-3978.

Almazán, J., Gordo, A., Fornés, A., & Valveny, E. (2014b). Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence, 36*(12), 2552-2566.

Alsharif, O., & Pineau, J. (2013). End-to-end text recognition with hybrid HMM maxout models. *arXiv preprint arXiv:1310.1811*.

Amrouch, M., & Rabi, M. (2017). *Deep Neural Networks Features for Arabic Handwriting Recognition.* Paper presented at the International Conference on Advanced Information Technology, Services and Systems.

Anbarjafari, G., & Demirel, H. (2010). Image super resolution based on interpolation of wavelet domain high frequency subbands and the spatial domain input image. *ETRI journal, 32*(3), 390-394.

Anbarjafari, G., Izadpanahi, S., & Demirel, H. (2015). Video resolution enhancement by using discrete and stationary wavelet transforms with illumination compensation. *Signal, Image and Video Processing, 9*(1), 87-92.

Anthimopoulos, M., Gatos, B., & Pratikakis, I. (2013). Detection of artificial and scene text in images and video frames. *Pattern Analysis and Applications, 16*(3), 431-446.

Ayyalasomayajula, K. R., & Brun, A. (2014). *Document binarization using topological clustering guided laplacian energy segmentation.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on.

Babaguchi, N., Yamada, K., Kise, K., & Tezuka, Y. (1991). Connectionist model binarization *Neural Networks In Pattern Recognition And Their Applications* (pp. 127-142): World Scientific.

Baker, S., & Kanade, T. (2002). Limits on super-resolution and how to break them. *IEEE transactions on pattern analysis and machine intelligence, 24*(9), 1167-1183.

Banerjee, P., Bhattacharya, U., & Chaudhuri, B. B. (2014). *Automatic detection of handwritten texts from video frames of lectures.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on.

Baran, R., Partila, P., & Wilk, R. (2018). *Automated Text Detection and Character Recognition in Natural Scenes Based on Local Image Features and Contour Processing Techniques.* Paper presented at the International Conference on Intelligent Human Systems Integration.

Barnsley, M. F. (2014). *Fractals everywhere*: Academic press.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). *Surf: Speeded up robust features.* Paper presented at the European conference on computer vision.

Ben-Ami, I., Basha, T., & Avidan, S. (2012). *Racing Bib Numbers Recognition.* Paper presented at the BMVC.

Ben-Ezra, M., Lin, Z., & Wilburn, B. (2007). *Penrose pixels super-resolution in the detector layout domain.* Paper presented at the Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.

Bernsen, J. (1986). *Dynamic thresholding of grey-level images.* Paper presented at the International conference on pattern recognition.

Bhardwaj, D., & Pankajakshan, V. (2016). Image Overlay Text Detection Based on JPEG Truncation Error Analysis. *IEEE Signal Processing Letters, 23*(8), 1027-1031.

Bhunia, A. K., Kumar, G., Roy, P. P., Balasubramanian, R., & Pal, U. (2017). Text recognition in scene image and video frame using Color Channel selection. *Multimedia Tools and Applications*, 1-28.

Biswas, B., Bhattacharya, U., & Chaudhuri, B. B. (2014). *A global-to-local approach to binarization of degraded document images.* Paper presented at the Pattern Recognition (ICPR), 2014 22nd International Conference on.

Bolan, S., Shijian, L., & Tan, C. L. (2010). *A self-training learning document binarization framework.* Paper presented at the Pattern Recognition (ICPR), 2010 20th International Conference on.

Bosamiya, J. H., Agrawal, P., Roy, P. P., & Balasubramanian, R. (2015). *Script independent scene text segmentation using fast stroke width transform and GrabCut.* Paper presented at the Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on.

Bosch, A., Zisserman, A., & Muñoz, X. (2008). Scene classification using a hybrid generative/discriminative approach. *IEEE transactions on pattern analysis and machine intelligence, 30*(4), 712-727.

Brezeale, D., & Cook, D. J. (2006). *Using closed captions and visual features to classify movies by genre.* Paper presented at the Poster session of the Seventh International Workshop on Multimedia Data Mining (MDM/KDD2006).

Buades, A., Coll, B., & Morel, J.-M. (2005). *A non-local algorithm for image denoising.* Paper presented at the Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.

Cai, M., Song, J., & Lyu, M. R. (2002). *A new approach for video text detection.* Paper presented at the Image Processing. 2002. Proceedings. 2002 International Conference on.

Capel, D., & Zisserman, A. (2000). *Super-resolution enhancement of text image sequences.* Paper presented at the Pattern Recognition, 2000. Proceedings. 15th International Conference on.

Chamchong, R., & Fung, C. C. (2010). *Optimal selection of binarization techniques for the processing of ancient palm leaf manuscripts.* Paper presented at the Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on.

Chang, H., Yeung, D.-Y., & Xiong, Y. (2004). *Super-resolution through neighbor embedding.* Paper presented at the Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on.

Chattopadhyay, T., Reddy, V. R., & Garain, U. (2013a). *Automatic selection of binarization method for robust OCR.* Paper presented at the 2013 12th International Conference on Document Analysis and Recognition.

Chattopadhyay, T., Reddy, V. R., & Garain, U. (2013b). *Automatic selection of binarization method for robust OCR.* Paper presented at the Document Analysis and Recognition (ICDAR), 2013 12th International Conference on.

Chen, C., Zhang, L., Bu, J., Wang, C., & Chen, W. (2010). Constrained Laplacian Eigenmap for dimensionality reduction. *Neurocomputing, 73*(4-6), 951-958.

Chen, D., & Odobez, J.-M. (2005). Video text recognition using sequential Monte Carlo and error voting methods. *Pattern Recognition Letters, 26*(9), 1386-1403.

Chen, D., Odobez, J.-M., & Bourlard, H. (2004). Text detection and recognition in images and video frames. *Pattern Recognition, 37*(3), 595-608.

Chen, H., Tsai, S. S., Schroth, G., Chen, D. M., Grzeszczuk, R., & Girod, B. (2011). *Robust text detection in natural images with edge-enhanced maximally stable extremal regions.* Paper presented at the Image Processing (ICIP), 2011 18th IEEE International Conference on.

Chen, X., & Gupta, A. (2015). *Webly supervised learning of convolutional networks.* Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

Chen, Y.-L., & Wu, B.-F. (2009). A multi-plane approach for text segmentation of complex document images. *Pattern Recognition, 42*(7), 1419-1444.

Chen, Y., & Wang, L. (2017). Broken and degraded document images binarization. *Neurocomputing, 237*, 272-280.

Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., & Zhou, S. (2017). *Focusing Attention: Towards Accurate Text Recognition in Natural Images.* Paper presented at the 2017 IEEE International Conference on Computer Vision (ICCV).

Cho, H., Sung, M., & Jun, B. (2016). *Canny text detector: Fast and robust scene text localization algorithm.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Chou, C.-H., Huang, C.-C., Lin, W.-H., & Chang, F. (2005). *Learning to binarize document images using a decision cascade.* Paper presented at the IEEE International Conference on Image Processing 2005.

Chou, C.-H., Lin, W.-H., & Chang, F. (2010). A binarization method with learning-built rules for document images produced by cameras. *Pattern Recognition, 43*(4), 1518-1530.

Cloud Vision, A. (2011). Derive insight from images with our powerful Cloud Vision API.

Coates, A., Carpenter, B., Case, C., Satheesh, S., Suresh, B., Wang, T., . . . Ng, A. Y. (2011). *Text detection and character recognition in scene images with unsupervised feature learning.* Paper presented at the Document Analysis and Recognition (ICDAR), 2011 International Conference on.

Conaire, C. O., O'Connor, N. E., & Smeaton, A. F. (2007). *Detector adaptation by maximising agreement between independent data sources.* Paper presented at the Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Crandall, D., Antani, S., & Kasturi, R. (2003). Extraction of special effects caption text events from digital video. *International journal on document analysis and recognition, 5*(2-3), 138-157.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). *Visual categorization with bags of keypoints.* Paper presented at the Workshop on statistical learning in computer vision, ECCV.

da Silva, L. F., Conci, A., & Sanchez, A. (2009). *Automatic discrimination between printed and handwritten text in documents.* Paper presented at the Computer

Graphics and Image Processing (SIBGRAPI), 2009 XXII Brazilian Symposium on.

Dalal, N., & Triggs, B. (2005). *Histograms of oriented gradients for human detection.* Paper presented at the Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.

Dalley, G., Freeman, B., & Marks, J. (2004). *Single-frame text super-resolution: A bayesian approach.* Paper presented at the Image Processing, 2004. ICIP'04. 2004 International Conference on.

Demirel, H., & Anbarjafari, G. (2011). Image resolution enhancement by using discrete and stationary wavelet decomposition. *IEEE transactions on image processing, 20*(5), 1458-1460.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *Imagenet: A large-scale hierarchical image database.* Paper presented at the Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.

Dimitrova, N., Agnihotri, L., & Wei, G. (2000). *Video classification based on HMM using text and faces.* Paper presented at the Signal Processing Conference, 2000 10th European.

Dlagnekov, L., & Belongie, S. J. (2005). *Recognizing cars*: Department of Computer Science and Engineering, University of California, San Diego.

Dobez, J.-M., & Chen, D. (2002). *Robust video text segmentation and recognition with multiple hypotheses.* Paper presented at the Image Processing. 2002. Proceedings. 2002 International Conference on.

Doermann, D., Liang, J., & Li, H. (2003). *Progress in camera-based document image analysis.* Paper presented at the Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). *Long-term recurrent convolutional networks for visual recognition and description.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Donaldson, K., & Myers, G. K. (2005). Bayesian super-resolution of text in videowith a text-specific bimodal prior. *International Journal of Document Analysis and Recognition (IJDAR), 7*(2-3), 159-167.

Dong, C., Loy, C. C., He, K., & Tang, X. (2014). *Learning a deep convolutional network for image super-resolution.* Paper presented at the European Conference on Computer Vision.

Dos Santos, J. E. B., Dubuisson, B., & Bortolozzi, F. (2002). *Characterizing and distinguishing text in bank cheque images.* Paper presented at the Computer Graphics and Image Processing, 2002. Proceedings. XV Brazilian Symposium on.

Dunlop, H. (2010). *Scene classification of images and video via semantic segmentation.* Paper presented at the Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on.

Ebrahimi, M., & Vrscay, E. R. (2007). *Solving the inverse problem of image zooming using "self-examples".* Paper presented at the International Conference Image Analysis and Recognition.

El-Yacoubi, A., Gilloux, M., Sabourin, R., & Suen, C. Y. (1999). An HMM-based approach for off-line unconstrained handwritten word modeling and recognition. *IEEE transactions on pattern analysis and machine intelligence, 21*(8), 752-760.

Elagouni, K., Garcia, C., Mamalet, F., & Sébillot, P. (2014). Text recognition in multimedia documents: a study of two neural-based ocrs using and avoiding character segmentation. *International Journal on Document Analysis and Recognition (IJDAR), 17*(1), 19-31.

Engine, G. A. (2008). CLOUD VISION API.

Epshtein, B., Ofek, E., & Wexler, Y. (2010). *Detecting text in natural scenes with stroke width transform.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on.

Evangelopoulos, G., Zlatintsi, A., Skoumas, G., Rapantzikos, K., Potamianos, A., Maragos, P., & Avrithis, Y. (2009). *Video event detection and summarization using audio, visual and text saliency.* Paper presented at the Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.

Farahmand, A., Sarrafzadeh, H., & Shanbehzadeh, J. (2017). Noise removal and binarization of scanned document images using clustering of features.

Farooq, F., Sridharan, K., & Govindaraju, V. (2006). *Identifying Handwritten Text in Mixed Documents.* Paper presented at the ICPR (2).

Fattal, R. (2007). *Image upsampling via imposed edge statistics.* Paper presented at the ACM transactions on graphics (TOG).

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition.

Fernando, B., Gavves, S., Mogrovejo, O., Antonio, J., Ghodrati, A., & Tuytelaars, T. (2015). *Modeling video evolution for action recognition.* Paper presented at the Proceedings CVPR 2015.

Fonseca, M. J., & Jorge, J. A. (2000). *Using fuzzy logic to recognize geometric shapes interactively.* Paper presented at the Fuzzy Systems, 2000. FUZZ IEEE 2000. The Ninth IEEE International Conference on.

Freeman, W. T., Jones, T. R., & Pasztor, E. C. (2002). Example-based super-resolution. *IEEE Computer graphics and Applications, 22*(2), 56-65.

Freeman, W. T., Pasztor, E. C., & Carmichael, O. T. (2000). Learning low-level vision. *International Journal of Computer Vision, 40*(1), 25-47.

Garcia, C., & Apostolidis, X. (2000). *Text detection and segmentation in complex color images.* Paper presented at the Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on.

Ghanei, S., & Faez, K. (2016). Localizing scene texts by fuzzy inference systems and low rank matrix recovery model. *Computer Vision and Image Understanding, 142*, 94-110.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). *Rich feature hierarchies for accurate object detection and semantic segmentation.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Gllavata, J., Ewerth, R., Stefi, T., & Freisleben, B. (2004). *Unsupervised text segmentation using color and wavelet features.* Paper presented at the International Conference on Image and Video Retrieval.

Gomez, L., & Karatzas, D. (2013). *Multi-script text extraction from natural scenes.* Paper presented at the Document Analysis and Recognition (ICDAR), 2013 12th International Conference on.

Gómez, L., & Karatzas, D. (2014). *MSER-based real-time text detection and tracking.* Paper presented at the Pattern Recognition (ICPR), 2014 22nd International Conference on.

Gonzalez, A., Bergasa, L. M., & Yebes, J. J. (2014). Text detection and recognition on traffic panels from street-level imagery using visual appearance. *IEEE Transactions on Intelligent Transportation Systems, 15*(1), 228-238.

Gonzalez, R. C., & Woods, R. E. (2002). Digital image processing: Prentice hall New Jersey.

Grafmüller, M., & Beyerer, J. (2013). Performance improvement of character recognition in industrial applications using prior knowledge for more reliable segmentation. *Expert Systems with Applications, 40*(17), 6955-6963.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). *Speech recognition with deep recurrent neural networks.* Paper presented at the Acoustics, speech and signal processing (icassp), 2013 ieee international conference on.

Greenhalgh, J., & Mirmehdi, M. (2015). Recognizing text-based traffic signs. *IEEE Transactions on Intelligent Transportation Systems, 16*(3), 1360-1369.

Guichard, F., & Morel, J.-M. (2003). A note on two classical enhancement filters and their associated pde's. *International Journal of Computer Vision, 52*(2-3), 153-160.

Guo, J. K., & Ma, M. Y. (2001). *Separating handwritten material from machine printed text using hidden markov models.* Paper presented at the Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on.

Han, X., Singh, B., Morariu, V. I., & Davis, L. S. (2017). VRFP: On-the-fly video retrieval using web images and fast fisher vector products. *IEEE Transactions on multimedia, 19*(7), 1583-1595.

He, J., Do, Q., Downton, A. C., & Kim, J. (2005). *A comparison of binarization methods for historical archive documents.* Paper presented at the Eighth International Conference on Document Analysis and Recognition (ICDAR'05).

He, P., Huang, W., Qiao, Y., Loy, C. C., & Tang, X. (2016). *Reading Scene Text in Deep Convolutional Sequences.* Paper presented at the AAAI.

Hooda, A., Kathuria, M., & Pankajakshan, V. (2014). *Application of forgery localization in overlay text detection.* Paper presented at the Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing.

Hou, H., & Andrews, H. (1978). Cubic splines for image interpolation and digital filtering. *IEEE Transactions on acoustics, speech, and signal processing, 26*(6), 508-517.

Howe, N. R. (2011). *A laplacian energy for document binarization.* Paper presented at the Document Analysis and Recognition (ICDAR), 2011 International Conference on.

Howe, N. R. (2013). Document binarization with automatic parameter tuning. *International Journal on Document Analysis and Recognition (IJDAR), 16*(3), 247-258.

Hua, X.-S., Wenyin, L., & Zhang, H.-J. (2004). An automatic performance evaluation protocol for video text detection algorithms. *IEEE Transactions on circuits and systems for video technology, 14*(4), 498-507.

Hua, X.-S., Yin, P., & Zhang, H.-J. (2002). *Efficient video text recognition using multiple frame integration.* Paper presented at the Image Processing. 2002. Proceedings. 2002 International Conference on.

Huang, Q., Liu, Z., Rosenberg, A., Gibbon, D., & Shahraray, B. (1999). *Automated generation of news content hierarchy by integrating audio, video, and text information.* Paper presented at the Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on.

Huang, R., Shivakumara, P., Yaokai, F., & Uchida, S. (2013). Scene character detection and recognition with cooperative multiple-hypothesis framework. *IEICE TRANSACTIONS on Information and Systems, 96*(10), 2235-2244.

Huang, W., Lin, Z., Yang, J., & Wang, J. (2013). *Text localization in natural images using stroke feature transform and text covariance descriptors.* Paper presented at the Computer Vision (ICCV), 2013 IEEE International Conference on.

Huang, W., Qiao, Y., & Tang, X. (2014). *Robust scene text detection with convolution neural network induced mser trees.* Paper presented at the European Conference on Computer Vision.

Huang, W., Shivakumara, P., & Tan, C. L. (2008). *Detecting moving text in video using temporal information.* Paper presented at the Pattern Recognition, 2008. ICPR 2008. 19th International Conference on.

Irani, M., & Peleg, S. (1991). Improving resolution by image registration. *CVGIP: Graphical models and image processing, 53*(3), 231-239.

Irani, M., & Peleg, S. (1993). Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation, 4*(4), 324-335.

Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.

Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2016). Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision, 116*(1), 1-20.

Jaderberg, M., Vedaldi, A., & Zisserman, A. (2014). *Deep features for text spotting.* Paper presented at the European conference on computer vision.

Jagannathan, L., & Jawahar, C. (2005). *Perspective correction methods for camera based document analysis.* Paper presented at the Proc. First Int. Workshop on Camera-based Document Analysis and Recognition.

Jain, A. K., & Bhattacharjee, S. (1992). Text segmentation using Gabor filters for automatic document processing. *Machine Vision and Applications, 5*(3), 169-184.

Jain, M., Mathew, M., & Jawahar, C. (2017). *Unconstrained scene text and video text recognition for Arabic script.* Paper presented at the Arabic Script Analysis and Recognition (ASAR), 2017 1st International Workshop on.

Jalab, H. A., & Ibrahim, R. W. (2015). Fractional Alexander polynomials for image denoising. *Signal Processing, 107*, 340-354.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence, 35*(1), 221-231.

Jia, F., Shi, C., He, K., Wang, C., & Xiao, B. (2016). *Document image binarization using structural symmetry of strokes.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on.

Jia, F., Shi, C., He, K., Wang, C., & Xiao, B. (2018). Degraded document image binarization using structural symmetry of strokes. *Pattern Recognition, 74*, 225-240.

Jianchao, Y., Wright, J., Huang, T., & Ma, Y. (2008). *Image super-resolution as sparse representation of raw image patches.* Paper presented at the Proc. IEEE Conf. on Computer Vision and Pattern Recognition.

Jung, K., Kim, K. I., & Jain, A. K. (2004). Text information extraction in images and video: a survey. *Pattern Recognition, 37*(5), 977-997.

Kakumanu, P., Makrogiannis, S., & Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. *Pattern Recognition, 40*(3), 1106-1122.

Kang, L., Li, Y., & Doermann, D. (2014). *Orientation robust text line detection in natural images.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., . . . Lu, S. (2015). *ICDAR 2015 competition on robust reading.* Paper presented at the Document Analysis and Recognition (ICDAR), 2015 13th International Conference on.

Karatzas, D., Mestre, S. R., Mas, J., Nourbakhsh, F., & Roy, P. P. (2011). *ICDAR 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email).* Paper presented at the Document Analysis and Recognition (ICDAR), 2011 International Conference on.

Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L. G., Mestre, S. R., . . . De Las Heras, L. P. (2013). *ICDAR 2013 robust reading competition.* Paper presented at the Document Analysis and Recognition (ICDAR), 2013 12th International Conference on.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). *Large-scale video classification with convolutional neural networks.* Paper presented at the Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.

Kavallieratou, E., & Stamatatos, S. (2004). *Discrimination of machine-printed from handwritten text using simple structural characteristics.* Paper presented at the Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on.

Keren, D., Peleg, S., & Brada, R. (1988). *Image sequence enhancement using sub-pixel displacements.* Paper presented at the Computer Vision and Pattern Recognition, 1988. Proceedings CVPR'88., Computer Society Conference on.

Kesidis, A. L., & Gatos, B. (2011). *Efficient cut-off threshold estimation for word spotting applications.* Paper presented at the Document Analysis and Recognition (ICDAR), 2011 International Conference on.

Khare, V., Shivakumara, P., Paramesran, R., & Blumenstein, M. (2017). Arbitrarily-oriented multi-lingual text detection in video. *Multimedia Tools and Applications, 76*(15), 16625-16655.

Khare, V., Shivakumara, P., & Raveendran, P. (2015). A new Histogram Oriented Moments descriptor for multi-oriented moving text detection in video. *Expert Systems with Applications, 42*(21), 7627-7640.

Khare, V., Shivakumara, P., Raveendran, P., & Blumenstein, M. (2016). A blind deconvolution model for scene text detection and recognition in video. *Pattern Recognition, 54*, 128-148.

Khare, V., Shivakumara, P., Raveendran, P., Meng, L. K., & Woon, H. H. (2015). *A new sharpness based approach for character segmentation in License plate images.* Paper presented at the Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on.

Kim, K. I., Jung, K., & Kim, J. H. (2003). Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE transactions on pattern analysis and machine intelligence, 25*(12), 1631-1639.

Kim, W., & Kim, C. (2009). A new approach for overlay text detection and extraction from complex video scene. *IEEE transactions on image processing, 18*(2), 401-411.

Klaser, A., Marszałek, M., & Schmid, C. (2008). *A spatio-temporal descriptor based on 3d-gradients.* Paper presented at the BMVC 2008-19th British Machine Vision Conference.

Kopf, S., Haenselmann, T., & Effelsberg, W. (2005). *Robust character recognition in low-resolution images and videos*: Universität Mannheim/Institut für Informatik.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Paper presented at the Advances in neural information processing systems.

Kuk, J. G., Cho, N. I., & Lee, K. M. (2008). *MAP-MRF approach for binarization of degraded document image.* Paper presented at the Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on.

Kumar, D., Prasad, M., & Ramakrishnan, A. (2012). Benchmarking recognition results on word image datasets. *arXiv preprint arXiv:1208.6137*.

Kumar, G., & Govindaraju, V. (2017). Bayesian background models for keyword spotting in handwritten documents. *Pattern Recognition, 64*, 84-91.

Kumari, L., Dey, V., & Raheja, J. (2018). A Three-Layer Approach for Overlay Text Extraction in Video Stream *Soft Computing: Theories and Applications* (pp. 79-87): Springer.

Lamba, V. (2008). *Neuro fuzzy systems*: University Science Press.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision, 64*(2-3), 107-123.

Laskin, N. (2003). Fractional poisson process. *Communications in Nonlinear Science and Numerical Simulation, 8*(3-4), 201-213.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories.* Paper presented at the Computer vision and pattern recognition, 2006 IEEE computer society conference on.

Lazzara, G., & Géraud, T. (2014). Efficient multiscale Sauvola's binarization. *International Journal on Document Analysis and Recognition (IJDAR), 17*(2), 105-123.

Lee, C.-Y., Bhardwaj, A., Di, W., Jagadeesh, V., & Piramuthu, R. (2014). *Region-based discriminative feature pooling for scene text recognition.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Lee, S. J., & Kim, S. W. (2016). *Recognition of Slab Identification Numbers Using a Deep Convolutional Neural Network.* Paper presented at the Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on.

Lelore, T., & Bouchara, F. (2009). *Document image binarisation using markov field model.* Paper presented at the Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on.

Li, H., & Doermann, D. (1999). *Text enhancement in digital video using multiple frame integration.* Paper presented at the Proceedings of the seventh ACM international conference on Multimedia (Part 1).

Li, H., & Doermann, D. (2000). *Superresolution-based enhancement of text in digital video.* Paper presented at the Pattern Recognition, 2000. Proceedings. 15th International Conference on.

Li, H., Doermann, D., & Kia, O. (2000). Automatic text detection and tracking in digital video. *IEEE transactions on image processing, 9*(1), 147-156.

Li, H., Kia, O. E., & Doermann, D. S. (1999). *Text enhancement in digital video.* Paper presented at the Electronic Imaging'99.

Li, H., Luo, W., & Huang, J. (2015). Anti-forensics of double JPEG compression with the same quantization matrix. *Multimedia Tools and Applications, 74*(17), 6729-6744.

Li, M., Zhang, X., & Mao, J. (2008). Neighboring region variance weighted mean image fusion based on wavelet transform. *Foreign Electronic Measurement Technology, 27*(1), 5-6.

Li, X., & Orchard, M. T. (2001). New edge-directed interpolation. *IEEE transactions on image processing, 10*(10), 1521-1527.

Liang, G., Shivakumara, P., Lu, T., & Tan, C. L. (2015a). Multi-spectral fusion based approach for arbitrarily oriented scene text detection in video images. *IEEE transactions on image processing, 24*(11), 4488-4501.

Liang, G., Shivakumara, P., Lu, T., & Tan, C. L. (2015b). *A new wavelet-Laplacian method for arbitrarily-oriented character segmentation in video text lines.* Paper presented at the Document Analysis and Recognition (ICDAR), 2015 13th International Conference on.

Liao, M., Shi, B., Bai, X., Wang, X., & Liu, W. (2017). *TextBoxes: A Fast Text Detector with a Single Deep Neural Network.* Paper presented at the AAAI.

Lienhart, R., & Maydt, J. (2002). *An extended set of haar-like features for rapid object detection.* Paper presented at the Image Processing. 2002. Proceedings. 2002 International Conference on.

Lienhart, R., & Wernicke, A. (2002). Localizing and segmenting text in images and videos. *IEEE Transactions on circuits and systems for video technology, 12*(4), 256-268.

Lin, Z., & Shum, H.-Y. (2004). Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE transactions on pattern analysis and machine intelligence, 26*(1), 83-97.

Liu, C., Wang, C., & Dai, R. (2005). *Text detection in images based on unsupervised classification of edge-based features.* Paper presented at the Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on.

Liu, D., Lai, K.-T., Ye, G., Chen, M.-S., & Chang, S.-F. (2013). *Sample-specific late fusion for visual category recognition.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on.

Liu, J., Chen, C., Zhu, Y., Liu, W., & Metaxas, D. N. (2016). Video classification via weakly supervised sequence modeling. *Computer Vision and Image Understanding, 152*, 79-87.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234*, 11-26.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110.

Lu, S., & Tan, C.-L. (2007). *Keyword spotting and retrieval of document images captured by a digital camera.* Paper presented at the Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on.

Lu, Y., & Tan, C. L. (2002). *Word spotting in Chinese document images without layout analysis.* Paper presented at the Pattern Recognition, 2002. Proceedings. 16th International Conference on.

Lucas, S. M., Panaretos, A., Sosa, L., Tang, A., Wong, S., & Young, R. (2003). *ICDAR 2003 robust reading competitions.* Paper presented at the Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on.

Lucchese, L., & Cortelazzo, G. M. (2000). A noise-robust frequency domain technique for estimating planar roto-translations. *IEEE Transactions on Signal Processing, 48*(6), 1769-1786.

Lyu, M. R., Song, J., & Cai, M. (2005). A comprehensive method for multilingual video text detection, localization, and extraction. *IEEE Transactions on circuits and systems for video technology, 15*(2), 243-255.

Mallat, S. (1999). *A wavelet tour of signal processing*: Academic press.

Mallat, S., & Yu, G. (2010). Super-resolution with sparse mixing estimators. *IEEE transactions on image processing, 19*(11), 2889-2900.

Mancas-Thillou, C., & Gosselin, B. (2007). Color text extraction with selective metric-based clustering. *Computer Vision and Image Understanding, 107*(1), 97-107.

Mancas-Thillou, C., & Mirmehdi, M. (2005). *Super-resolution text using the teager filter.* Paper presented at the First International Workshop on Camera-Based Document Analysis and Recognition.

Manisha, S., & Sharmila, T. S. (2016). *Text frame classification and recognition using segmentation technique.* Paper presented at the Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on.

Michaeli, T., & Irani, M. (2013). *Nonparametric blind super-resolution.* Paper presented at the Computer Vision (ICCV), 2013 IEEE International Conference on.

Milyaev, S., Barinova, O., Novikova, T., Kohli, P., & Lempitsky, V. (2013). *Image binarization for end-to-end text understanding in natural images.* Paper presented at the Document Analysis and Recognition (ICDAR), 2013 12th International Conference on.

Milyaev, S., Barinova, O., Novikova, T., Kohli, P., & Lempitsky, V. (2015). Fast and accurate scene text understanding with image binarization and off-the-shelf OCR. *International Journal on Document Analysis and Recognition (IJDAR), 18*(2), 169-182.

Mishra, A., Alahari, K., & Jawahar, C. (2011). *An MRF model for binarization of natural scene text.* Paper presented at the 2011 International Conference on Document Analysis and Recognition.

Mishra, A., Alahari, K., & Jawahar, C. (2012a). *Scene text recognition using higher order language priors.* Paper presented at the BMVC 2012-23rd British Machine Vision Conference.

Mishra, A., Alahari, K., & Jawahar, C. (2012b). *Top-down and bottom-up cues for scene text recognition.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.

Mishra, A., Alahari, K., & Jawahar, C. (2017). Unsupervised refinement of color and stroke features for text binarization. *International Journal on Document Analysis and Recognition (IJDAR), 20*(2), 105-121.

Mithe, R., Indalkar, S., & Divekar, N. (2013). Optical character recognition. *International Journal of Recent Technology and Engineering, 2*(1), 72-75.

Mittal, A., Roy, P. P., Singh, P., & Raman, B. (2017). Rotation and script independent text detection from video frames using sub pixel mapping. *Journal of Visual Communication and Image Representation, 46*, 187-198.

Moghaddam, R. F., & Cheriet, M. (2010). A multi-scale framework for adaptive binarization of degraded document images. *Pattern Recognition, 43*(6), 2186-2198.

Mondal, T., Ragot, N., Ramel, J.-Y., & Pal, U. (2016). Flexible Sequence Matching technique: An effective learning-free approach for word spotting. *Pattern Recognition, 60*, 596-612.

Mori, G., & Malik, J. (2003). *Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA.* Paper presented at the Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on.

Mosleh, A., Bouguila, N., & Hamza, A. B. (2013). Automatic inpainting scheme for video text detection and removal. *IEEE transactions on image processing, 22*(11), 4460-4472.

Navon, Y. (2008). *Layer-based binarization for textual images.* Paper presented at the Pattern Recognition, 2008. ICPR 2008. 19th International Conference on.

Ng, J. Y.-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). *Beyond short snippets: Deep networks for video classification.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). *Multimodal deep learning.* Paper presented at the Proceedings of the 28th international conference on machine learning (ICML-11).

Nguyen, P. X., Wang, K., & Belongie, S. (2014). *Video text detection and recognition: Dataset and benchmark.* Paper presented at the Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on.

Niblack, W. (1985). *An introduction to digital image processing*: Strandberg Publishing Company.

Nogueira, K., Penatti, O. A., & dos Santos, J. A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition, 61*, 539-556.

Novikova, T., Barinova, O., Kohli, P., & Lempitsky, V. (2012). *Large-lexicon attribute-consistent text recognition in natural images.* Paper presented at the European Conference on Computer Vision.

Ntirogiannis, K., Gatos, B., & Pratikakis, I. (2011). *Binarization of textual content in video frames.* Paper presented at the 2011 International Conference on Document Analysis and Recognition.

Ntirogiannis, K., Gatos, B., & Pratikakis, I. (2013). Performance evaluation methodology for historical document image binarization. *IEEE transactions on image processing, 22*(2), 595-609.

Ohn-Bar, E., & Trivedi, M. M. (2017). Are all objects equal? Deep spatio-temporal importance prediction in driving videos. *Pattern Recognition, 64*, 425-436.

Oneata, D., Verbeek, J., & Schmid, C. (2013). *Action and event recognition with fisher vectors on a compact feature set.* Paper presented at the Computer Vision (ICCV), 2013 IEEE International Conference on.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). *Learning and transferring mid-level image representations using convolutional neural networks.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on.

Osher, S., & Rudin, L. I. (1990). Feature-oriented image enhancement using shock filters. *SIAM Journal on numerical analysis, 27*(4), 919-940.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics, 9*(1), 62-66.

Pal, U., & Chaudhuri, B. (1999). *Automatic separation of machine-printed and hand-written text lines.* Paper presented at the Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on.

Pal, U., & Chaudhuri, B. (2001). *Automatic identification of english, chinese, arabic, devnagari and bangla script line.* Paper presented at the Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on.

Pal, U., & Chaudhuri, B. B. (2001). Machine-printed and hand-written text lines identification. *Pattern Recognition Letters, 22*(3-4), 431-441.

Pal, U., Roy, P. P., Tripathy, N., & Lladós, J. (2010). Multi-oriented Bangla and Devnagari text recognition. *Pattern Recognition, 43*(12), 4124-4136.

Park, J.-G., & Kim, K.-J. (2013). Design of a visual perception model with edge-adaptive Gabor filter and support vector machine for traffic sign detection. *Expert Systems with Applications, 40*(9), 3679-3687.

Park, J., Kwon, Y., & Kim, J. H. (2005). *An example-based prior model for text image super-resolution.* Paper presented at the Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on.

Pawlak, Z., & Skowron, A. (2007). Rudiments of rough sets. *Information sciences, 177*(1), 3-27.

Peng, X., Setlur, S., Govindaraju, V., & Sitaram, R. (2013). Handwritten text separation from annotated machine printed documents using Markov Random Fields. *International Journal on Document Analysis and Recognition (IJDAR), 16*(1), 1-16.

Perronnin, F., Sánchez, J., & Mensink, T. (2010). *Improving the fisher kernel for large-scale image classification.* Paper presented at the European conference on computer vision.

Phan, T. Q., Shivakumara, P., & Tan, C. L. (2009). *A Laplacian method for video text detection.* Paper presented at the Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on.

Phan, T. Q., Shivakumara, P., Tian, S., & Tan, C. L. (2013). *Recognizing text with perspective distortion in natural scenes.* Paper presented at the Computer Vision (ICCV), 2013 IEEE International Conference on.

Piao, Y., & Park, H. (2007). *Image resolution enhancement using inter-subband correlation in wavelet domain.* Paper presented at the Image Processing, 2007. ICIP 2007. IEEE International Conference on.

Podlubny, I. (1998). *Fractional differential equations: an introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications* (Vol. 198): Elsevier.

Polesel, A., Ramponi, G., & Mathews, V. J. (2000). Image enhancement via adaptive unsharp masking. *IEEE transactions on image processing, 9*(3), 505-510.

Pratikakis, I., Gatos, B., & Ntirogiannis, K. (2012). *ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012).* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on.

Puigcerver, J., Toselli, A. H., & Vidal, E. (2015). *Probabilistic interpretation and improvements to the hmm-filler for handwritten keyword spotting.* Paper presented at the Document Analysis and Recognition (ICDAR), 2015 13th International Conference on.

Qi, W., Gu, L., Jiang, H., Chen, X.-R., & Zhang, H.-J. (2000). *Integrating visual, audio and text analysis for news video.* Paper presented at the Image Processing, 2000. Proceedings. 2000 International Conference on.

Qin, L., Shivakumara, P., Lu, T., Pal, U., & Tan, C. L. (2016). *Video scene text frames categorization for text detection and recognition.* Paper presented at the Pattern Recognition (ICPR), 2016 23rd International Conference on.

Rabi, M., Amrouch, M., & Mahani, Z. (2018). Recognition of Cursive Arabic Handwritten Text Using Embedded Training Based on Hidden Markov Models.

International Journal of Pattern Recognition and Artificial Intelligence, 32(01), 1860007.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). *Objects in context.* Paper presented at the Computer vision, 2007. ICCV 2007. IEEE 11th international conference on.

Raghunandan, K., Shivakumara, P., Kumar, G. H., Pal, U., & Lu, T. (2016). *New Sharpness Features for Image Type Classification Based on Textual Information.* Paper presented at the Document Analysis Systems (DAS), 2016 12th IAPR Workshop on.

Rathore, M. M., Ahmad, A., Paul, A., & Rho, S. (2016). Urban planning and building smart cities based on the internet of things using big data analytics. *Computer Networks, 101*, 63-80.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). *CNN features off-the-shelf: an astounding baseline for recognition.* Paper presented at the Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on.

Reddy, B. S., & Chatterji, B. N. (1996). An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE transactions on image processing, 5*(8), 1266-1271.

Retsinas, G., Louloudis, G., Stamatopoulos, N., & Gatos, B. (2016). *Keyword spotting in handwritten documents using projections of oriented gradients.* Paper presented at the Document Analysis Systems (DAS), 2016 12th IAPR Workshop on.

Riba, P., Lladós, J., Fornés, A., & Dutta, A. (2015). *Large-scale graph indexing using binary embeddings of node contexts.* Paper presented at the International Workshop on Graph-Based Representations in Pattern Recognition.

Risnumawan, A., Shivakumara, P., Chan, C. S., & Tan, C. L. (2014). A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications, 41*(18), 8027-8048.

Rong, L., Suyu, W., & Shi, Z. (2014). *A two level algorithm for text detection in natural scene images.* Paper presented at the Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on.

Roy, P. P., Bhunia, A. K., Bhattacharyya, A., & Pal, U. (2017). Word Searching in Scene Image and Video Frame in Multi-Script Scenario using Dynamic Shape Coding. *arXiv preprint arXiv:1708.05529*.

Roy, P. P., Bhunia, A. K., & Pal, U. (2018). Date-field retrieval in scene image and video frames using text enhancement and shape coding. *Neurocomputing, 274*, 37-49.

Roy, P. P., Pal, U., Lladós, J., & Delalandre, M. (2012). Multi-oriented touching text character segmentation in graphical documents using dynamic programming. *Pattern Recognition, 45*(5), 1972-1983.

Roy, S., Roy, P. P., Shivakumara, P., Louloudis, G., Tan, C. L., & Pal, U. (2013a). *HMM-based multi oriented text recognition in natural scene image.* Paper presented at the Pattern Recognition (ACPR), 2013 2nd IAPR Asian Conference on.

Roy, S., Roy, P. P., Shivakumara, P., Louloudis, G., Tan, C. L., & Pal, U. (2013b). *HMM-based multi oriented text recognition in natural scene image.* Paper presented at the 2013 2nd IAPR Asian Conference on Pattern Recognition.

Roy, S., Roy, P. P., Shivakumara, P., & Pal, U. (2013). *Word recognition in natural scene and video images using Hidden Markov Model.* Paper presented at the Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on.

Roy, S., Shivakumara, P., Pal, U., Lu, T., & Tan, C. L. (2016). *New Tampered Features for Scene and Caption Text Classification in Video Frame.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on.

Roy, S., Shivakumara, P., Roy, P. P., Pal, U., Tan, C. L., & Lu, T. (2015). Bayesian classifier for multi-oriented video text recognition system. *Expert Systems with Applications, 42*(13), 5554-5566.

Roy, S., Shivakumara, P., Roy, P. P., & Tan, C. L. (2012). *Wavelet-gradient-fusion for video text binarization.* Paper presented at the Pattern Recognition (ICPR), 2012 21st International Conference on.

Rudin, L. I. (1987). Images, numerical analysis of singularities and shock filters.

Saidane, Z., & Garcia, C. (2007). *Robust binarization for video text recognition.* Paper presented at the Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on.

Sain, A., Bhunia, A. K., Roy, P. P., & Pal, U. (2018). Multi-oriented text detection and verification in video frames and scene images. *Neurocomputing, 275*, 1531-1549.

Sari, T., Kefali, A., & Bahi, H. (2012). *An MLP for binarizing images of old manuscripts.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on.

Sattar, F., & Tay, D. B. (1999). Enhancement of document images using multiresolution and fuzzy logic techniques. *IEEE Signal Processing Letters, 6*(10), 249-252.

Sauvola, J., Seppanen, T., Haapakoski, S., & Pietikainen, M. (1997). *Adaptive document binarization.* Paper presented at the Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on.

Schavemaker, J. G., Reinders, M. J., Gerbrands, J. J., & Backer, E. (2000). Image sharpening by morphological filtering. *Pattern Recognition, 33*(6), 997-1012.

Sehad, A., Chibani, Y., & Cheriet, M. (2014). *Gabor filters for degraded document image binarization.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on.

Shahab, A., Shafait, F., & Dengel, A. (2011). *ICDAR 2011 robust reading competition challenge 2: Reading text in scene images.* Paper presented at the Document Analysis and Recognition (ICDAR), 2011 International Conference on.

Shahraray, B., & Gibbon, D. C. (1995). *Automated authoring of hypermedia documents of video programs.* Paper presented at the Proceedings of the third ACM international conference on Multimedia.

Sharma, A. (2015). *Adapting off-the-shelf cnns for word spotting & recognition.* Paper presented at the Document Analysis and Recognition (ICDAR), 2015 13th International Conference on.

Sharma, N., Pal, U., & Blumenstein, M. (2012). *Recent advances in video based document processing: a review.* Paper presented at the Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on.

Sharma, N., Shivakumara, P., Pal, U., Blumenstein, M., & Tan, C. L. (2012). *A new method for arbitrarily-oriented text detection in video.* Paper presented at the Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on.

Sharma, N., Shivakumara, P., Pal, U., Blumenstein, M., & Tan, C. L. (2015). Piece-wise linearity based method for text frame classification in video. *Pattern Recognition, 48*(3), 862-881.

Shi, B., Bai, X., & Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence, 39*(11), 2298-2304.

Shi, B., Wang, X., Lyu, P., Yao, C., & Bai, X. (2016). *Robust scene text recognition with automatic rectification.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Shi, C., Ruan, Q., & An, G. (2014). Sparse feature selection based on graph Laplacian for web image annotation. *Image and Vision Computing, 32*(3), 189-201.

Shi, C., Xiao, B., Wang, C., & Zhang, Y. (2012). *Adaptive graph cut based binarization of video text images.* Paper presented at the Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on.

Shivakumara, P., Dutta, A., Phan, T. Q., Tan, C. L., & Pal, U. (2011). A novel mutual nearest neighbor based symmetry for text frame classification in video. *Pattern Recognition, 44*(8), 1671-1683.

Shivakumara, P., Dutta, A., Tan, C. L., & Pal, U. (2014). Multi-oriented scene text detection in video based on wavelet and angle projection boundary growing. *Multimedia Tools and Applications, 72*(1), 515-539.

Shivakumara, P., Huang, W., Phan, T. Q., & Tan, C. L. (2010). Accurate video text detection through classification of low and high contrast images. *Pattern Recognition, 43*(6), 2165-2185.

Shivakumara, P., Kumar, N. V., Guru, D., & Tan, C. (2014). *Separation of graphics (superimposed) and scene text in video frames.* Paper presented at the Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on.

Shivakumara, P., Liang, G., Roy, S., Pal, U., & Lu, T. (2015). *New texture-spatial features for keyword spotting in video images.* Paper presented at the Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on.

Shivakumara, P., Phan, T. Q., Lu, S., & Tan, C. L. (2013). Gradient vector flow and grouping-based method for arbitrarily oriented scene text detection in video images. *IEEE Transactions on circuits and systems for video technology, 23*(10), 1729-1739.

Shivakumara, P., Phan, T. Q., & Tan, C. L. (2009). *Video text detection based on filters and edge features.* Paper presented at the Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on.

Shivakumara, P., Phan, T. Q., & Tan, C. L. (2010). New Fourier-statistical features in RGB space for video text detection. *IEEE Transactions on circuits and systems for video technology, 20*(11), 1520-1532.

Shivakumara, P., Phan, T. Q., & Tan, C. L. (2010). *New wavelet and color features for text detection in video.* Paper presented at the Pattern Recognition (ICPR), 2010 20th International Conference on.

Shivakumara, P., Phan, T. Q., & Tan, C. L. (2011). A laplacian approach to multi-oriented text detection in video. *IEEE transactions on pattern analysis and machine intelligence, 33*(2), 412-419.

Shivakumara, P., Raghavendra, R., Qin, L., Raja, K. B., Lu, T., & Pal, U. (2017). A new multi-modal approach to bib number/text detection and recognition in Marathon images. *Pattern Recognition, 61*, 479-491.

Shivakumara, P., Sreedhar, R. P., Phan, T. Q., Lu, S., & Tan, C. L. (2012). Multioriented video scene text detection through Bayesian classification and boundary growing. *IEEE Transactions on circuits and systems for video technology, 22*(8), 1227-1235.

Shivakumara, P., Suhil, M., Guru, D., & Tan, C. L. (2014). *A new Laplacian method for arbitrarily-oriented word segmentation in video.* Paper presented at the Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on.

Shivakumara, P., & Tan, C. L. (2010). *Novel edge features for text frame classification in video.* Paper presented at the Pattern Recognition (ICPR), 2010 20th International Conference on.

Shivakumara, P., Wu, L., Lu, T., Tan, C. L., Blumenstein, M., & Anami, B. S. (2017). Fractals based multi-oriented text detection system for recognition in mobile video images. *Pattern Recognition, 68*, 158-174.

Simonyan, K., & Zisserman, A. (2014). *Two-stream convolutional networks for action recognition in videos.* Paper presented at the Advances in neural information processing systems.

Singh, A., & Ahuja, N. (2014). *Super-resolution using sub-band self-similarity.* Paper presented at the Asian Conference on Computer Vision.

Singh, A., Porikli, F., & Ahuja, N. (2014). *Super-resolving noisy images.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Singh, B. M., Sharma, R., Ghosh, D., & Mittal, A. (2014). Adaptive binarization of severely degraded and non-uniformly illuminated documents. *International Journal on Document Analysis and Recognition (IJDAR), 17*(4), 393-412.

Smith, R. (2007). *An overview of the Tesseract OCR engine.* Paper presented at the Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on.

Snoek, C. G., Worring, M., & Smeulders, A. W. (2005). *Early versus late fusion in semantic video analysis.* Paper presented at the Proceedings of the 13th annual ACM international conference on Multimedia.

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015). *Unsupervised learning of video representations using lstms.* Paper presented at the International conference on machine learning.

Su, B., & Lu, S. (2014). *Accurate scene text recognition based on recurrent neural network.* Paper presented at the Asian Conference on Computer Vision.

Su, B., Lu, S., & Tan, C. L. (2013). Robust document image binarization technique for degraded document images. *IEEE transactions on image processing, 22*(4), 1408-1417.

Sudholt, S., & Fink, G. A. (2016). *PHOCNet: A deep convolutional neural network for word spotting in handwritten documents.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on.

Sun, J., Zheng, N.-N., Tao, H., & Shum, H.-Y. (2003). *Image hallucination with primal sketch priors.* Paper presented at the Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on.

Sun, L., Huo, Q., Jia, W., & Chen, K. (2014). *Robust text detection in natural scene images by generalized color-enhanced contrasting extremal region and neural networks.* Paper presented at the Pattern Recognition (ICPR), 2014 22nd International Conference on.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks.* Paper presented at the Advances in neural information processing systems.

Tang, K., Fei-Fei, L., & Koller, D. (2012). *Learning latent temporal structure for complex event detection.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.

Tarafdar, A., Pal, U., Roy, P. P., Ragot, N., & Ramel, J.-Y. (2013). *A two-stage approach for word spotting in graphical documents.* Paper presented at the Document Analysis and Recognition (ICDAR), 2013 12th International Conference on.

Thévenaz, P., Blu, T., & Unser, M. (2000). Image interpolation and resampling. *Handbook of medical imaging, processing and analysis, 1*(1), 393-420.

Tian, D., Sun, H., & Vetro, A. (2016). *Keypoint trajectory coding on compact descriptor for video analysis.* Paper presented at the Image Processing (ICIP), 2016 IEEE International Conference on.

Tian, S., Pei, W.-Y., Zuo, Z.-Y., & Yin, X.-C. (2016). *Scene Text Detection in Video by Learning Locally and Globally.* Paper presented at the IJCAI.

Tian, S., Yin, X.-C., Su, Y., & Hao, H.-W. (2018). A unified framework for tracking based text detection and recognition from web videos. *IEEE transactions on pattern analysis and machine intelligence, 40*(3), 542-554.

Ting, K. M., Zhou, G.-T., Liu, F. T., & Tan, S. C. (2013). Mass estimation. *Machine learning, 90*(1), 127-160.

Toselli, A. H., Vidal, E., Romero, V., & Frinken, V. (2016). HMM word graph based keyword spotting in handwritten document images. *Information sciences, 370*, 497-518.

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., & Paluri, M. (2014). C3D: generic features for video analysis. *CoRR, abs/1412.0767, 2*(7), 8.

Tsai, R. (1984). Multiframe image restoration and registration. *Advance Computer Visual and Image Processing, 1*, 317-339.

Tu, Z., Chen, X., Yuille, A. L., & Zhu, S.-C. (2005). Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision, 63*(2), 113-140.

Vail, D. L., Veloso, M. M., & Lafferty, J. D. (2007). *Conditional random fields for activity recognition.* Paper presented at the Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems.

Vandewalle, P., Süsstrunk, S., & Vetterli, M. (2006). A frequency domain approach to registration of aliased images with application to super-resolution. *EURASIP journal on advances in signal processing, 2006*(1), 071459.

Wagdy, M., Faye, I., & Rohaya, D. (2015). Document image binarization using retinex and global thresholding. *ELCVIA Electronic Letters on Computer Vision and Image Analysis, 14*(1).

Wakahara, T., & Kita, K. (2011). *Binarization of color character strings in scene images using k-means clustering and support vector machines.* Paper presented at the Document Analysis and Recognition (ICDAR), 2011 International Conference on.

Wang, H., & Schmid, C. (2013). *Action recognition with improved trajectories.* Paper presented at the Computer Vision (ICCV), 2013 IEEE International Conference on.

Wang, K., Babenko, B., & Belongie, S. (2011a). *End-to-end scene text recognition.* Paper presented at the 2011 International Conference on Computer Vision.

Wang, K., Babenko, B., & Belongie, S. (2011b). *End-to-end scene text recognition.* Paper presented at the Computer Vision (ICCV), 2011 IEEE International Conference on.

Wang, K., & Belongie, S. (2010). *Word spotting in the wild.* Paper presented at the European Conference on Computer Vision.

Wang, L., Fan, W., He, Y., Sun, J., Katsuyama, Y., & Hotta, Y. (2014). *Fast and accurate text detection in natural scene images with user-intention.* Paper presented at the Pattern Recognition (ICPR), 2014 22nd International Conference on.

Wang, Q., Tang, X., & Shum, H. (2005). *Patch based blind image super resolution.* Paper presented at the Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on.

Wang, T., Wu, D. J., Coates, A., & Ng, A. Y. (2012). *End-to-end text recognition with convolutional neural networks.* Paper presented at the Pattern Recognition (ICPR), 2012 21st International Conference on.

Wang, W., Wu, Y., Shivakumara, P., & Lu, T. (2018). *Cloud of Line Distribution and Random Forest Based Text Detection from Natural/Video Scene Images.* Paper presented at the International Conference on Multimedia Modeling.

Wang, X., Farhadi, A., & Gupta, A. (2016). *Actions~ transformations.* Paper presented at the Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.

Wang, X., Jiang, Y., Yang, S., Zhu, X., Li, W., Fu, P., . . . Luo, Z. (2017). *End-to-End Scene Text Recognition in Videos Based on Multi Frame Tracking.* Paper presented at the Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on.

Wang, X., Song, Y., Zhang, Y., & Xin, J. (2015). Natural scene text detection with multi-layer segmentation and higher order conditional random field based analysis. *Pattern Recognition Letters, 60*, 41-47.

Wang, Y., & Mori, G. (2009). *Max-margin hidden conditional random fields for human action recognition.* Paper presented at the Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing, 13*(4), 600-612.

Wei, H., Gao, G., & Su, X. (2015). *A multiple instances approach to improving keyword spotting on historical Mongolian document images.* Paper presented at the Document Analysis and Recognition (ICDAR), 2015 13th International Conference on.

Wei, Y. C., & Lin, C. H. (2012). A robust video text detection approach using SVM. *Expert Systems with Applications, 39*(12), 10832-10840.

Weickert, J. (2003). *Coherence-enhancing shock filters.* Paper presented at the Joint Pattern Recognition Symposium.

Weinman, J. J., Learned-Miller, E., & Hanson, A. R. (2009). Scene text recognition using similarity and a lexicon with sparse belief propagation. *IEEE transactions on pattern analysis and machine intelligence, 31*(10), 1733-1746.

Wicht, B., Fischer, A., & Hennebert, J. (2016). *Deep learning features for handwritten keyword spotting.* Paper presented at the Pattern Recognition (ICPR), 2016 23rd International Conference on.

Wigington, C., Stewart, S., Davis, B., Barrett, B., Price, B., & Cohen, S. (2017). *Data Augmentation for Recognition of Handwritten Words and Lines Using a CNN-LSTM Network.* Paper presented at the Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on.

Wolf, C., Jolion, J.-M., & Chassaing, F. (2002). *Text localization, enhancement and binarization in multimedia documents.* Paper presented at the Pattern Recognition, 2002. Proceedings. 16th International Conference on.

Wong, E. K., & Chen, M. (2003). A new robust algorithm for video text extraction. *Pattern Recognition, 36*(6), 1397-1406.

Wu, L., Shivakumara, P., Lu, T., & Tan, C. L. (2014). *Text detection using delaunay triangulation in video sequence.* Paper presented at the Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on.

Wu, L., Shivakumara, P., Lu, T., & Tan, C. L. (2015). A new technique for multi-oriented scene text line detection and tracking in video. *IEEE Transactions on multimedia, 17*(8), 1137-1152.

Wu, Y., Shivakumara, P., Lu, T., Tan, C. L., Blumenstein, M., & Kumar, G. H. (2016). Contour restoration of text components for recognition in video/scene images. *IEEE transactions on image processing, 25*(12), 5622-5634.

Wu, Y., Wang, W., Palaiahnakote, S., & Lu, T. (2017). *A Robust Symmetry-based Method for Scene/Video Text Detection Through Neural Network.* Paper presented at the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR).

Wu, Z., Jiang, Y.-G., Wang, J., Pu, J., & Xue, X. (2014). *Exploring inter-feature and inter-class relationships with deep neural networks for video classification.* Paper presented at the Proceedings of the 22nd ACM international conference on Multimedia.

Wu, Z., Jiang, Y.-G., Wang, X., Ye, H., & Xue, X. (2016). *Multi-stream multi-class fusion of deep networks for video classification.* Paper presented at the Proceedings of the 2016 ACM on Multimedia Conference.

X. H. Yang, W. H., F. Yin and C. L. Liu. (2017). A Unified Video Text Detection Method with Network Flow. pp. 331-336

Xu, C., Wang, J., Wan, K., Li, Y., & Duan, L. (2006). *Live sports event detection based on broadcast video and web-casting text.* Paper presented at the Proceedings of the 14th ACM international conference on Multimedia.

Xu, C., Zhang, Y.-F., Zhu, G., Rui, Y., Lu, H., & Huang, Q. (2008). Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on multimedia, 10*(7), 1342-1355.

Xu, J., Shivakumara, P., Lu, T., Phan, T. Q., & Tan, C. L. (2014). *Graphics and scene text classification in video.* Paper presented at the Pattern Recognition (ICPR), 2014 22nd International Conference on.

Xu, J., Shivakumara, P., Lu, T., Tan, C. L., & Uchida, S. (2016). A new method for multi-oriented graphics-scene-3D text classification in video. *Pattern Recognition, 49*, 19-42.

Xu, Y., Shan, S., Qiu, Z., Jia, Z., Shen, Z., Wang, Y., . . . Chang, C. (2018). End-to-end subtitle detection and recognition for videos in East Asian languages via CNN ensemble. *Signal Processing: Image Communication, 60*, 131-143.

Xu, Z., Hu, J., & Deng, W. (2016). *Recurrent convolutional neural network for video classification.* Paper presented at the Multimedia and Expo (ICME), 2016 IEEE International Conference on.

Yang, C.-Y., Huang, J.-B., & Yang, M.-H. (2010). *Exploiting self-similarities for single frame super-resolution.* Paper presented at the Asian conference on computer vision.

Yang, C., Pei, W.-Y., Wu, L.-H., & Yin, X.-C. (2018). Chinese text-line detection from web videos with fully convolutional networks. *Big Data Analytics, 3*(1), 2.

Yang, C., Yin, X.-C., Li, Z., Wu, J., Guo, C., Wang, H., & Xiao, L. (2017). AdaDNNs: Adaptive Ensemble of Deep Neural Networks for Scene Text Recognition. *arXiv preprint arXiv:1710.03425*.

Yang, C., Yin, X.-C., Pei, W.-Y., Tian, S., Zuo, Z.-Y., Zhu, C., & Yan, J. (2017). Tracking Based Multi-Orientation Scene Text Detection: A Unified Framework With

Dynamic Programming. *IEEE transactions on image processing, 26*(7), 3235-3248.

Yang, J., Wang, Z., Lin, Z., Cohen, S., & Huang, T. (2012). Coupled dictionary training for image super-resolution. *IEEE transactions on image processing, 21*(8), 3467-3478.

Yang, Y., Song, J., Huang, Z., Ma, Z., Sebe, N., & Hauptmann, A. G. (2013). Multi-feature fusion via hierarchical regression for multimedia analysis. *IEEE Transactions on multimedia, 15*(3), 572-581.

Yao, C., Bai, X., & Liu, W. (2014). A unified framework for multioriented text detection and recognition. *IEEE transactions on image processing, 23*(11), 4737-4749.

Yao, C., Bai, X., Liu, W., Ma, Y., & Tu, Z. (2012). *Detecting texts of arbitrary orientations in natural images.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.

Ye, G., Liu, D., Jhuo, I.-H., & Chang, S.-F. (2012). *Robust late fusion with rank minimization.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.

Ye, P., & Doermann, D. (2013). *Document image quality assessment: A brief survey.* Paper presented at the Document Analysis and Recognition (ICDAR), 2013 12th International Conference on.

Ye, Q., & Doermann, D. (2015). Text detection and recognition in imagery: A survey. *IEEE transactions on pattern analysis and machine intelligence, 37*(7), 1480-1500.

Yi, C., & Tian, Y. (2012). Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification. *IEEE transactions on image processing, 21*(9), 4256-4268.

Yi, J., Peng, Y., & Xiao, J. (2009a). *Using multiple frame integration for the text recognition of video.* Paper presented at the 2009 10th International Conference on Document Analysis and Recognition.

Yi, J., Peng, Y., & Xiao, J. (2009b). *Using multiple frame integration for the text recognition of video.* Paper presented at the Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on.

Yin, X.-C., Pei, W.-Y., Zhang, J., & Hao, H.-W. (2015). Multi-orientation scene text detection with adaptive clustering. *IEEE transactions on pattern analysis and machine intelligence, 37*(9), 1930-1937.

Yin, X.-C., Yin, X., Huang, K., & Hao, H.-W. (2014). Robust text detection in natural scene images. *IEEE transactions on pattern analysis and machine intelligence, 36*(5), 970-983.

Yin, X.-C., Zuo, Z.-Y., Tian, S., & Liu, C.-L. (2016). Text detection, tracking and recognition in video: A comprehensive survey. *IEEE transactions on image processing, 25*(6), 2752-2773.

Yokobayashi, M., & Wakahara, T. (2005). *Segmentation and recognition of characters in scene images using selective binarization in color space and gat correlation.* Paper presented at the Eighth International Conference on Document Analysis and Recognition (ICDAR'05).

Yokobayashi, M., & Wakahara, T. (2006). *Binarization and recognition of degraded characters using a maximum separability axis in color space and gat correlation.* Paper presented at the 18th International Conference on Pattern Recognition (ICPR'06).

Young, S., Jansen, J., Odell, J., Ollason, D., & Woodland, P. (1995). The HTK Hidden Markov Model Toolkit Book, Entropic Cambridge Research Laboratory.

Yousfi, S., Berrani, S.-A., & Garcia, C. (2015). *Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos.* Paper presented at the Document Analysis and Recognition (ICDAR), 2015 13th International Conference on.

Yousfi, S., Berrani, S.-A., & Garcia, C. (2017). Contribution of recurrent connectionist language models in improving LSTM-based Arabic text recognition in videos. *Pattern Recognition, 64*, 245-254.

Yu, Y., Pedrycz, W., & Miao, D. (2013). Neighborhood rough sets based multi-label classification for automatic image annotation. *International Journal of Approximate Reasoning, 54*(9), 1373-1387.

Yuan, J., Wei, B., Liu, Y., Zhang, Y., & Wang, L. (2015). A method for text line detection in natural images. *Multimedia Tools and Applications, 74*(3), 859-884.

Yusufu, T., Wang, Y., & Fang, X. (2013). *A video text detection and tracking system.* Paper presented at the Multimedia (ISM), 2013 IEEE International Symposium on.

Zamberletti, A., Gallo, I., & Noce, L. (2015). *Augmented text character proposals and convolutional neural networks for text spotting from scene images.* Paper presented at the Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on.

Zeyde, R., Elad, M., & Protter, M. (2010). *On single image scale-up using sparse-representations.* Paper presented at the International conference on curves and surfaces.

Zhang, D., & Chang, S.-F. (2002). *Event detection in baseball video using superimposed caption recognition.* Paper presented at the Proceedings of the tenth ACM international conference on Multimedia.

Zhang, H., Zhao, K., Song, Y.-Z., & Guo, J. (2013). Text extraction from natural scene image: A survey. *Neurocomputing, 122*, 310-323.

Zhang, J., & Kasturi, R. (2008). *Extraction of text objects in video documents: Recent progress.* Paper presented at the 2008 The Eighth IAPR International Workshop on Document Analysis Systems.

Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision, 73*(2), 213-238.

Zhang, S., Liu, W., & Qin, Y. (2016). *Wake-up-word spotting using end-to-end deep neural network system.* Paper presented at the Pattern Recognition (ICPR), 2016 23rd International Conference on.

Zhang, X., Pal, U., & Tan, C. L. (2014). *Segmentation-free Keyword spotting for Bangla handwritten documents.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on.

Zhang, X., & Wu, X. (2008). Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation. *IEEE transactions on image processing, 17*(6), 887-896.

Zhang, Z., & Wang, W. (2013a). *A novel approach for binarization of overlay text.* Paper presented at the 2013 IEEE International Conference on Systems, Man, and Cybernetics.

Zhang, Z., & Wang, W. (2013b). *A novel approach for binarization of overlay text.* Paper presented at the Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on.

Zhang, Z., Wang, W., & Lu, K. (2014). *Video text extraction using the fusion of color gradient and Log-Gabor filter.* Paper presented at the Pattern Recognition (ICPR), 2014 22nd International Conference on.

Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., & Bai, X. (2016). Multi-oriented text detection with fully convolutional networks. *arXiv preprint arXiv:1604.04018*.

Zhao, G., & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence, 29*(6), 915-928.

Zhao, X., Lin, K.-H., Fu, Y., Hu, Y., Liu, Y., & Huang, T. S. (2011). Text from corners: a novel approach to detect text and caption in videos. *IEEE transactions on image processing, 20*(3), 790-799.

Zhao, Z., Fang, C., Lin, Z., & Wu, Y. (2015). A robust hybrid method for text detection in natural scenes by learning-based partial differential equations. *Neurocomputing, 168*, 23-34.

Zheng, Y., Li, H., & Doermann, D. (2004). Machine printed text and handwriting identification in noisy document images. *IEEE transactions on pattern analysis and machine intelligence, 26*(3), 337-353.

Zheng, Y., Li, Q., Liu, J., Liu, H., Li, G., & Zhang, S. (2017). A cascaded method for text detection in natural scene images. *Neurocomputing, 238*, 307-315.

Zhiwei, Z., Linlin, L., & Lim, T. C. (2010). *Edge based binarization for video text images.* Paper presented at the Pattern Recognition (ICPR), 2010 20th International Conference on.

Zhong, Y., Zhang, H., & Jain, A. K. (2000). Automatic caption localization in compressed video. *IEEE transactions on pattern analysis and machine intelligence, 22*(4), 385-392.

Zhou, C., Yan, T., Tao, W., & Lui, S. (2012). *A study of images denoising based on two improved fractional integral marks.* Paper presented at the International Conference on Intelligent Computing.

Zhou, J., Xu, L., Xiao, B., & Dai, R. (2007). *A robust system for text extraction in video.* Paper presented at the Machine Vision, 2007. ICMV 2007. International Conference on.

Zhou, M.-K., Zhang, X.-Y., Yin, F., & Liu, C.-L. (2016). Discriminative quadratic feature learning for handwritten Chinese character recognition. *Pattern Recognition, 49*, 7-18.

Zhu, A., Gao, R., & Uchida, S. (2016). Could scene context be beneficial for scene text detection? *Pattern Recognition, 58*, 204-215.

Zhu, Y., & Du, J. (2018). Sliding Line Point Regression for Shape Robust Scene Text Detection. *arXiv preprint arXiv:1801.09969*.

Zhu, Y., Yao, C., & Bai, X. (2016). Scene text detection and recognition: Recent advances and future trends. *Frontiers of Computer Science, 10*(1), 19-36.

Zuo, Z.-Y., Tian, S., Pei, W.-y., & Yin, X.-C. (2015). *Multi-strategy tracking based text detection in scene videos.* Paper presented at the Document Analysis and Recognition (ICDAR), 2015 13th International Conference on.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Journals:

Published:

[1] Roy, S., Shivakumara, P., Roy, P. P., Pal, U., Tan, C. L., & Lu, T. (2015). Bayesian classifier for multi-oriented video text recognition system. Expert Systems with Applications, 42(13), 5554-5566.

[2] Roy, S., Shivakumara, P., Jalab, H. A., Ibrahim, R. W., Pal, U., & Lu, T. (2016). Fractional Poisson enhancement model for text detection and recognition in video frames. Pattern Recognition, 52, 433-447.

[3] Roy, S., Shivakumara, P., Jain, N., Khare, V., Dutta, A., Pal, U., & Lu, T. (2018). Rough-Fuzzy based Scene Categorization for Text Detection and Recognition in Video. Pattern Recognition.

[4] K. S. Raghunandan, Palaiahnakote Shivakumara, Sangheeta Roy, G. Hemantha Kumar, Umapda Pal, Tong Lu, Member, IEEE, (2018, March). "Multi-Oriented-Type Text Detection and Recognition in Video/Scene/Born Digital Images," (accepted in IEEE Transactions on Circuits and Systems for Video Technology (CSVT)).

Under Review:

[1] Fractional Means based Method for Multi-Oriented Keyword Spotting in Video/Scene/License Plate Images.

Conferences:

[1] Roy, S., Shivakumara, P., Mondal, P., Raghavendra, R., Pal, U., & Lu, T. (2015, September). A New Multi-modal Technique for Bib Number/Text Detection in Natural Images. In Pacific Rim Conference on Multimedia (pp. 483-494). Springer, Cham.

[2] Shivakumara, P., Liang, G., Roy, S., Pal, U., & Lu, T. (2015, November). New texture-spatial features for keyword spotting in video images. In Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on (pp. 391-395). IEEE.

[3] Roy, S., Shivakumara, P., Pal, U., Lu, T., & Tan, C. L. (2016, October). New Tampered Features for Scene and Caption Text Classification in Video Frame. In Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on(pp. 36-41). IEEE.

[4] Roy, S., Shivakumara, P., Jain, N., Khare, V., Pal, U., & Lu, T. (2017, November). New Fuzzy-Mass Based Features for Video Image Type Categorization. In Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on (Vol. 1, pp. 838-843). IEEE.

[5] Roy, S., Shivakumara, P., Pal, U., Lu, T., & Wahab, A. W. B. A. (2017, November). Temporal Integration for Word-Wise Caption and Scene Text Identification. In Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on (Vol. 1, pp. 349-354). IEEE.