# LOCALIZATION AND RECOGNITION OF TEXT IN DIGITAL MEDIA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET SARACOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

November 2007

Approval of the Thesis

**"LOCALIZATION AND RECOGNITION OF TEXT IN DIGITAL MEDIA"**

Submitted by **AHMET SARACOĞLU** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering** by,

Prof. Dr. Canan Özgen

Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. İsmet Erkmen

Head of Department, **Electrical and Electronics Engineering** _____

Assoc. Prof. Dr. A. Aydın Alatan

Supervisor, **Electrical and Electronics Engineering, METU** _____

**Examining Committee Members**

Prof. Dr. Uğur Halıcı
Electrical and Electronics Engineering, METU _____

Assoc. Prof. Dr. A. Aydın Alatan
Electrical and Electronics Engineering, METU _____

Assoc. Prof. Dr. Gözde Bozdağı Akar
Electrical and Electronics Engineering, METU _____

Asst. Prof. Dr. Afşar Saranlı
Electrical and Electronics Engineering, METU _____

Asst. Prof. Dr. Pınar Duygulu Şahin
Computer Engineering, Bilkent _____

**Date: 27.11.2007**

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

**Name, Lastname : Ahmet Saracoğlu**

**Signature :**

# ABSTRACT

## LOCALIZATION AND RECOGNITION OF TEXT IN DIGITAL MEDIA

Saracoğlu, Ahmet
M.S., Department of Electrical and Electronics Engineering
Supervisor: Assoc. Prof. Dr. A. Aydın Alatan

November 2007, 113 pages

Textual information within digital media can be used in many areas such as, indexing and structuring of media databases, in the aid of visually impaired, translation of foreign signs and many more. This said, mainly text can be separated into two categories in digital media as, overlay-text and scene-text. In this thesis localization and recognition of video text regardless of its category in digital media is investigated. As a necessary first step, framework of a complete system is discussed. Next, a comparative analysis of feature vector and classification method pairs is presented. Furthermore, multi-part nature of text is exploited by proposing a novel Markov Random Field approach for the classification of text/non-text regions. Additionally, better localization of text is achieved by introducing bounding-box extraction method. And for the recognition of text regions, a handprint based Optical Character Recognition system is thoroughly investigated. During the investigation of text recognition, multi-hypothesis approach for the segmentation of background is proposed by incorporating k-Means clustering. Furthermore, a novel dictionary-based ranking mechanism is proposed for recognition spelling correction. And overall system is

simulated on a challenging data set. Also, a through survey on scene-text localization and recognition is presented. Furthermore, challenges are identified and discussed by providing related work on them. Scene-text localization simulations on a public competition data set are also provided. Lastly, in order to improve recognition performance of scene-text on signs that are affected from perspective projection distortion, a rectification method is proposed and simulated.

Keywords: overlay-text, scene-text, Video OCR, character recognition, Markov Random Fields, text localization, perspective rectification.

# ÖZ

## SAYISAL ORTAMDA BULUNAN YAZILARIN KONUMLANDIRILMASI VE TANINMASI

Saracoğlu, Ahmet
Yüksek Lisans, Elektrik-Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Doç. Dr. A. Aydın Alatan

Kasım 2007, 113 sayfa

Sayısal görüntü ortamlarında bulunan yazı bilgisi çok farklı alanlarda kullanılabilir örneğin; medya veritabanlarının dizinlenmesi ve yapılandırılmasında, görme engellilere yardımda, yabancı dillerdeki tabelaların çevirisinde ve daha bir çoğunda kullanılabilir. Sayısal ortamda bulunan yazılar yapay yazılar ve sahne yazıları olarak iki ana kategoriye ayrılabilir. Bu tezde, kategorisi ne olursa olsun sayısal ortamda bulunan yazıların yerlerinin bulunması ve tanınması üzerine araştırma yapılmıştır. İlk adım olarak bütün sistemin çerçevesi üzerinde durulmuştur. Daha sonra, öznitelik vektörleri ve sınıflandırma yöntemleri karşılaştırmalı olarak incelenmiştir. Ayrıca, yazının çok parçalı doğası Markov Rasgele Alanları yaklaşımı önerilerek değerlendirilmiştir. Buna ek olarak yazının daha iyi konumlandırılması için karakter kutusu çıkarma yöntemi öne sürülmüştür. Yazı alanlarının tanınması için ise elyazısına dayalı Optik Karakter Tanıma sistemi detaylı bir şekilde incelenmiştir. İnceleme sırasında, arka planın ayrılması için çoklu-hipotez yaklaşımına dayalı ve k-Ortalamalar yöntemini kullanan bir yöntem kullanılmıştır. Ayrıca tanıma sonuçlarının düzeltilmesi için sözlük tabanlı sıralama yöntemi önerilmiştir ve de sistemin tamamı güç bir veri kümesi üzerinde benzetimlenmiştir. Sahne yazılarınn bulunması ve tanınması

üzerinde detaylı bir araştırma da sunulmuştur. Bununla birlikte zorluklar belirlenmiş ve ilgili çalışmalar ele alınmıştır. Ayrıca, sahne yazılarınn konumlandırılması kamusal bir yarışma veri kümesi üzerinde benzetimlenmiştir. Son olarak da, levha üzerinde bulunan ve perspektif izdüşüm bozulumundan etkilenen yazıların doğrultulması için bir yöntem önerilmiş ve sınanmıştır.

Anahtar Kelimeler: yapay yazı, sahne yazısı, Video OCR, karakter tanıma, Markov Rasgele Alanları, yazı konumlandırma, perspektif doğrultma.

To Mom and Dad,

# ACKNOWLEDGEMENTS

I would like to express my gratitude and deep appreciation to my supervisor Assoc. Prof. Dr. A. Aydın Alatan for his guidance, positive suggestions and also for the great research environment he had provided.

I would like to also express my thanks for their great friendship and assistance to Oytun Akman, Cevahir Çığla and Yoldaş Ataseven. We were together for two invaluable years and I will surely miss working with them.

I would like to thank my friends in Multimedia Research Group and Space Technologies Research Institute for such friendly research environments they had provided. I have learned much from their suggestions and experiences.

I would like to also acknowledge The Scientific and Technological Research Council of Turkey (TUBITAK) for their funds.

Finally, I would like to thank my Mom and Dad for their love, support and patience over the years. This thesis is dedicated to them.

# TABLE OF CONTENTS

# LIST OF TABLES

TABLES

# LIST OF FIGURES

FIGURES

# CHAPTER 1

# INTRODUCTION

Classification of audio-visual perception plays a vital role in our lives. As we understand speech, read newspaper or distinguish between friend and foe, we utilize this brilliant tool. Although, we perform these feats and many more facilely, underlying mechanisms are exceptionally complex. In this sense, pattern recognition can be described as the process of assigning a category to a given data and acting according to this assignment. Furthermore, in today's world, indispensability of machines in the aid of humans is apparent. Moreover, it is not a surprise that there are countless systems employing pattern recognition principles. Optical character recognition, speech recognition, data mining and face recognition are just a few examples from myriad application areas of such systems.

## *1.1 Problem Definition and Application Areas*

One of the earliest applications of data classification for information extraction is Optical Character Recognition (OCR). Very briefly, OCR technology converts images of printed or handwritten text documents into machine-editable text. Moreover, it has been proven to be beneficial in many situations, e.g., libraries and museums incorporate OCR technology for their digital archives; post offices employ this technology to increase the degree of automation present in mail processing facilities. An ambitious extension to OCR technology would be the

recognition of text from any digital media – generic images and/or videos – not just from document images; otherwise, stated the concept of Video OCR.

Mainly, text in digital media can be separated into two groups, as *"overlay-text"* and *"scene-text"* (Figure 1). While former one represents the text that is artificially added, superimposed into the video frames, and much easier to detect, the latter one refers to the text which already exists in the scene. Regardless of the category of text, extracting this information from digital media would be advantageous in many areas.


**(a)** **(b)**

**Figure 1: Example images of scene-text (a) and overlay text (b).**

Such information, for example, can be used to index any video database quite efficiently and effectively. Speaker information, location, date/time, score results of sport games, etc. can be queried more thoroughly, since these information can be determined as text data in the digital media. Moreover, it should be noted that, recognition of text is easier compared to the extraction/recognition of objects in a complex scene due to the simple fact that text is inherently designed to be readable.

Furthermore, the quality of life of visually impaired can be improved extremely by designing a system that can recognize text from a simple restaurant menu on the fly. Even better, extracting information from traffic signs, posters, notice boards and etc. would be a huge leap for the visually impaired people. Moreover, text recognition from signs and boards can also be used for translation purposes. Such a translation system would further be useful, when dealing with languages, whose characters are difficult to understand and copy directly. As a final application, spam e-mail analysis could be performed better via this technology, due to the fact it is becoming more common to distribute advertorial information within attached images, instead of plain text in the body of the e-mail.

Although the aforementioned benefits of the embedded text as an information source are indisputable, still there is no dependable and robust method to extract textual information from the visual content.

In this thesis, fundamental building blocks of text information extraction system are investigated. Intuitively, these building blocks can be decomposed into three main branches; namely, *Text Localization Module*, *Text Recognition Module* and *Post-Correction Module.* The initial module focuses on determining the locations of text regions within digital media; the latter one recognizes the underlying characters in these regions and the final block uses various techniques for correcting any errors observed in the text results.

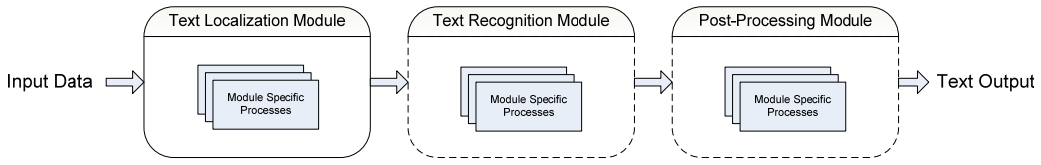



**Figure 2: Overview of Text Information Extraction System.**

## *1.2 Thesis Outline*

The dissertation is composed of two main parts each addressing the efforts devoted to different categories of text.

In Chapter 2, *Overlay-Text Detection and Recognition* is investigated thoroughly. A complete system, capable of both locating the text, as well as recognizing the characters within each detected text block is proposed and tested with extensive simulations. This chapter starts by presenting the related work on the literature. Afterwards, each building block for localization and recognition is discussed with comparative analysis of the proposed methods.

Chapter 3 is devoted on the *Scene-Text Detection and Recognition* problem. The major challenges of the problem are identified in the beginning of this chapter. Related work on general localization methods coupled with literature survey focusing on challenges are also reported. Localization performance is demonstrated on a public data set. Furthermore, in order to improve the recognition performance perspective distortion compensation method is proposed for scene-text present on signs.

Finally, in Chapter 4, the summary of the thesis is presented. Moreover, concluding remarks and future work are discussed in detail.

# CHAPTER 2

# OVERLAY-TEXT DETECTION AND RECOGNITION

Although the benefits of an overlay-text recognition system are obvious, there are still some challenges that worth mentioning. An overlay-text can be positioned arbitrarily in a generic video frame. Besides, text instances in a frame can be in any color and/or font. Moreover, uniform background assumption could be quite weak. Lastly, but not the least, video frames may suffer from noise or low-resolution and even from the motion artifacts within text regions.

In this chapter, overlay-text detection and recognition problem is examined thoroughly. In the first part, a detailed analysis of an overlay-text localization system is discussed. In the following section, a recognition module is described in detail.

## *2.1 Overlay-Text Localization*

Localization can be summarized as the process of discovering the positions of text regions in a video frame, or an image, and determining bounding boxes around these text regions for further processing. By the help of this stage, the amount of data to be processed is also minimized; thus, the computational burden and complexity of the subsequent stages is decreased. It should be noted that in order to recognize the text in an effective manner, extra care should be taken for

localization step, since any error in the localization stage would propagate through the succeeding stages.

### *2.1.1 Related Work on Overlay-Text Localization*

Although in some of the previous studies [1], [17] and [18], the proposed methods for the solution of the overlay-text localization problem are divided into two main groups, as *connected-component-based* and *texture-based* approaches; a more appropriate classification could be obtained as follows: *gradient/edge-based methods, color-based methods* and *texture-based methods.* The gradient/edge-based methods generally assume that text characters preserve strong edges against the background regions. On the other hand, texture-based approaches hypothesize that text regions are indeed special textures and the methods in this group primarily differ from each other in terms of the representation of this text texture. Finally, in color-based methods, the overlay-text regions are detected by assuming the uniformity of color intensity of the existing text within frames.

#### *2.1.1.1 Gradient/Edge-based Methods*

In [7], Smith and Kanade first apply a 3×3 horizontal differential filter on the entire frame with a binary thresholding for extraction of vertical edge features. Afterwards, dilation-based smoothing operation is employed in order to connect adjacent edges, i.e. character regions, while removing small edge fragments. After determining edge clusters, heuristic based elimination is applied to the clusters to further purge non-text regions. These heuristic parameters are size, fill factor and aspect ratio. Finally, intensity histogram is used to test for high contrast nature of overlay-text. However, this method lacks of a decision step other than plain thresholds and heuristics in order to classify text/non-text regions.

As a similar approach to [9], Wolf et al. propose a detection method which exploits the fact that text characters form a regular texture containing vertical strokes [10]. Their method is established on a previous work of LeBourgeois [11]

which locates text with a measure of accumulated gradients. Wolf et al. firstly extract horizontal edges by horizontal Sobel operator. After this step, binarization (text/non-text) coupled with erosion and dilation operations are applied. Lastly, some geometrical constraints are employed on the extracted rectangles. This approach, although shows acceptable detection rates, cannot handle high number of false alarms produced from the binarization of the accumulated gradients and morphological operations.

In [17] and [16], text detection and extraction is carried out by edge detection, local thresholding, and hysteresis edge recovery. Afterwards, a coarse-to-fine localization scheme is performed to identify text regions precisely. In their recent work [16], Lyu et al. introduce a modified Sobel edge detector in which alterations are made to favor perpendicular corners. Moreover, in that work sequential multi-resolution is proposed to detect large font text regions. Although promising results are demonstrated throughout the study, depending exclusively on edge extraction, the heuristics for text detection would yield a weak detection scheme in terms of localization. Moreover, bounding box extraction method proposed in the approach cannot handle complex background.

#### *2.1.1.2 Color-based Methods*

Zhong, et al. [4] locate text regions based on color segmentation and horizontal spatial variance. In their proposed method, candidate text lines are extracted by using horizontal spatial variance on the grayscale version of image. The horizontal spatial variance is obtained by employing a Canny edge detector. After finding connected components of edge pixels, some heuristic rules are used to reduce possible non-text regions in a final step. In the second part, under the assumption of uniform text color, further connected component analysis is employed. By adopting a hybrid approach, both color segmentation and spatial variance analysis are used in conjunction, the precision of the overall system is tried to be improved. However, this method still relies on mostly crude thresholds

and heuristics for the classification problem, slightly underestimating text-like edges present in generic images.

In another approach [9], color uniformity and horizontal alignment assumptions are utilized. Text strokes are extracted by using a color clustering method on rows of the image. In their method, perceptually uniform color space *L*a*b** is chosen and the text regions are merged by using heuristics on the lengths of short strokes, as well as gaps and height. However, these heuristics may not capture the nature of lowercase characters, since base and top lines of a lowercase text would not be aligned. Moreover, as stated in the paper this method cannot detect non-uniform colored text regions.

In [6], Dimitrova et al. detects caption text by initially assuming that caption text has yellow or white color and then the method seeks strong edges in red and green color channels. After some refinements, the text regions are verified by employing the temporal redundancy of text in video. These assumptions might not be valid for more generic text other than subtitles and caption text. In [3] and [14], Lienhart present yet another color clustering with both spatial and temporal heuristics for detecting the text in title and credits of a video.

#### *2.1.1.3 Texture-based Methods*

In [1], Li et al. present a supervised learning-based method for detection and a sum of squared difference (SSD) based method for tracking overlay-text in digital video. In their localization method initially, 16×16 windows are used to compute feature vectors and afterwards each feature vector are classified as text or non-text by using a neural network. They calculate features from the wavelet transform sub-bands by computing the mean, second-order and third order central moments. More detailed information on their supervised training method could be found in [2]. Additionally in [2], skew detection is also introduced by estimating the moments of connected components.

Shin et al. [12] propose an SVM-based text detection method in which for detection of various text sizes, a pyramid of images approach is used. Moreover, the gray values of a local window are directly fed to SVM without incorporating any feature extraction method. However interestingly, not all of the pixel values are used. Instead some of them are selected by a star shaped map. In their follow-up study [13], the text regions are identified by employing continuously adaptive mean shift (CAMSHIFT) algorithm on the SVM output. Accordingly, a rectangular search window, which represents text regions, is adapted continuously to fit the probability distribution within the window. In each iteration position, simply width and height of the search window is controlled. Although this method has an acceptable miss-rate performance overall false detection rate suffers from the data representation scheme adopted in the method.

In [5], Chaddha et al. propose a Discrete Cosine Transform (DCT) based method among others to detect text regions in JPEG compressed images. In their DCT-based approach, a set of absolute value of DCT coefficients are summed and this sum is compared with a predetermined threshold to classify text regions. Crandall et al. further improves previous method by using only a subset of DCT coefficients calculated from 8×8 blocks in [8]. This subset of coefficients is chosen by examining the absolute values of text and non-text coefficients. The decision of text or non-text is given by using thresholding on computed texture energies which by the way obtained by summing absolute values of coefficients. Afterwards, further localization step is employed, in which connected components of text blocks are refined into bounding rectangles by an iterative greedy algorithm. In a previous study [18], Zhong et al. also use a subset of DCT coefficients for modeling the texture of text-like structures. From their analysis on DCT coefficients, they conclude that horizontal and vertical harmonics best represent the properties of text lines. Thus, they detect and refine text regions by calculating horizontal and vertical text energies, which are computed by summing mid-band horizontal and vertical harmonics. Unfortunately, previous approaches

lack a classifier at the discrimination step, which decreases the potential and power of the whole system, since these methods rely solely on pre-computed thresholds.

Gllavata et al. in [15] incorporate a wavelet transform and unsupervised classification based method for detecting overlay-text in digital video. Firstly, wavelet transform is applied on the image. Assuming that high-frequency coefficients of wavelet transform represents text areas, an unsupervised clustering method, specifically *K-means* [32], is employed on the feature vector obtained from coefficients. The feature vectors are constructed from the standard deviations of high-frequency coefficients in a sliding window which moved over the transformed image. Further improvement is achieved by extracting bounding boxes by connected component analysis and horizontal projection profile. However, the initialization of cluster centers in the study is still an investigation point left in the paper.

Tseng et al. [19] gave an approach based on the fusion of different classification outputs. They combine detection results of previous works presented in [20] and [21]. In the first methodology, texture and motion energy features are used to generate candidate text regions. For each candidate region, color layering is used to generate several hypothetical binary images. The second method extracts candidate regions that may contain text characters. By separating each candidate character region, the technique employs successive refinement steps to preserve character regions. Proposed method in [19] combines the detection results of both techniques by using a normalized ensemble fusion method in order to obtain absolute detection result.

### *2.1.2 Framework for Overlay-Text Localization*

From the discussion on the previous studies, the following framework can be developed, as illustrated in Figure 3. The framework consists of three phases;

*feature extraction, classification* and *bounding-box extraction*. Before progressing further, it should be strictly remembered that text detection is a two-class problem with the following class labels; *text* and *non-text*.

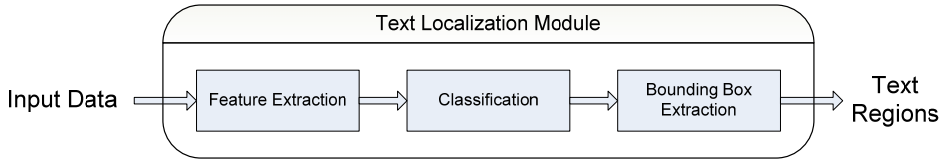



**Figure 3: Overlay-Text Localization Framework**

All of the prior works somehow strive to represent text as good as they can; some of them try exploiting color as a feature, whereas others use edge information. Although there are different approaches, the primary underlying reason is the same; if a proper representation can be determined, then classification problem between text and non-text would be trivial. Moreover, the dimensionality of the data is also reduced in this representation step. Assuming color uniformity as in the previously mentioned body of work, however plausible as it may seem, can yield to a weak overall system for text detection, because of complex color schemes on text and background, small sized text and even color degradation present on captured data. Furthermore, building text detection largely on edges would multiply the number of false results, since cityscape images contain a lot of text-like features; additionally, superimposed graphics, such as information bars on news videos and etc., can exhibit text-like features.

In this thesis, the smallest unit, for which text or non-text decision is performed, is selected as 8×8 image block, thus feature vectors are computed based on these units. After obtaining a feature vector, the decision between text and non-text is obtained. For the classification of pixels or regions as text, unfortunately most of the previous related works rely on heuristics and empirically calculated thresholds. Although this approach decreases the computational complexity as

well as the processing time, still using a mature classification technique would be far more superior. Finally, the classification results are further processed to obtain relatively tight bounding boxes around text regions for obtaining better text recognition performance. Although only a fraction of the previous studies investigated the text bounding-box extraction, it is a necessary and decisive step if a fully functional Video OCR is endeavored.

### *2.1.3 Feature Extraction for Overlay-Text Detection*

#### *2.1.3.1 Discrete Cosine Transform*

Discrete Cosine Transform (DCT) is the most widely used transform among a wide class of image coding systems, namely *transform coders*. This transform is used in JPEG compression standard, as well as in MPEG, MJPEG and DV video compression standards. This said, in our approach, instead of using partial energy of DCT coefficients, as in [5],[8],[18], the coefficients are directly utilized as feature vectors. At this point, one should note that the overlay-text in video frames have a strong contrasting characteristic with respect to their background. In other words, the text regions possess quite abrupt changes from the non-text regions (i.e. background), and these changes correspond to the high frequencies for DCT coefficients. Thus, for the sake of decreasing dimensionality, in this thesis not all 64 coefficients, but only the 27 coefficients, incurred from the mid- and high-indexed coefficients as in Figure 5, are utilized to obtain a feature vector due to the fact that coefficients for the highest DCT frequencies are usually distorted during compression, whereas low frequencies do not contain texture characteristics. This argument could be understood better by the example in Figure 4, in which the reconstruction of a frame by using only the coefficients in the band shown in Figure 5 is depicted.


**Figure 4: Example capture frame and reconstruction of the frame from selected DCT coefficients.**

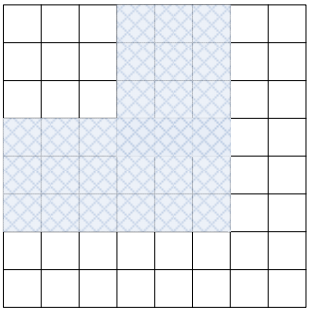


**Figure 5: Selected DCT Coefficients.**

### *2.1.3.2 Haar Wavelet Transform*

The property of wavelet transforms for yielding a finer resolution at the higher frequencies for detection of the edges make wavelet transforms a good candidate for feature extraction. By examining different scales wavelet transforms not only capture the frequency content but also the locations at which these frequencies are observed.

In addition to the ability of Haar Wavelet to capture text-like structures, such as lines, there are further underlying reasons for choosing Haar Wavelet over other wavelet transforms. Since this transform can be computed by the following simple

equations, the computational complexity of the transform is quite low, indeed $O(n)$.

$$LL_{x,y} = \frac{1}{4}\left(i_{2x,2y} + i_{2x,2y+1} + i_{2x+1,2y} + i_{2x+1,2y+1}\right) \quad (2.1)$$

$$LH_{x,y} = \frac{1}{4}\left(i_{2x,2y} - i_{2x,2y+1} + i_{2x+1,2y} - i_{2x+1,2y+1}\right) \quad (2.2)$$

$$HL_{x,y} = \frac{1}{4}\left(i_{2x,2y} + i_{2x,2y+1} - i_{2x+1,2y} - i_{2x+1,2y+1}\right) \quad (2.3)$$

$$HH_{x,y} = \frac{1}{4}\left(i_{2x,2y} - i_{2x,2y+1} - i_{2x+1,2y} + i_{2x+1,2y+1}\right) \quad (2.4)$$

In the set of equations $i_{x,y}$ represents image values at position $(x, y)$ and *LL, LH, HL, HH* abbreviations represent sub-bands.

Feature vector, however, constructed from *mean*, *second central moment* and *third central moment* of the blocks in the transform domain, as proposed in [1]. It should be noted that since 8×8 blocks are used only two levels of decomposition is useful for calculating aforementioned features. Thus, a 24-dimensional feature vector is obtained. Moreover, since some of the sub-bands – specifically *LH* and *HL* sub-bands – have more representation power of text structures over other sub-bands, the dimension of this vector can be decreased to 12 by only using features from these sub-bands.

### *2.1.3.3 MPEG-7 Edge Histogram Descriptor*

As one of the standardized MPEG-7 descriptors [22],[23],[24], edge histogram descriptor captures the distribution of certain edge types in an image. As shown in Figure 6, there are 5 edge types in the edge histogram descriptor. Since overlay-text contains edges similar to those considered in the descriptor, it has been hypothesized that this descriptor would yield a good representation performance.

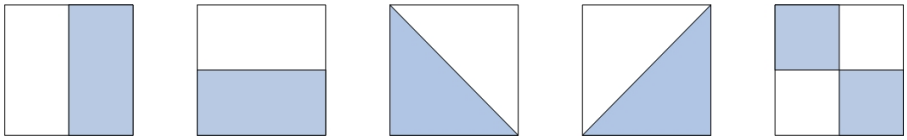


**Figure 6: (from left)Vertical, Horizontal, Diagonal, Off-Diagonal and Non-Directional Edges**

### *2.1.3.4 MPEG-7 Homogeneous Texture Descriptor*

The MPEG-7 homogeneous texture descriptor (HTD), as described in [22] and [25], consists of the mean and the standard deviation of an image, energy and energy deviation of Fourier transform of the image. The feature values – means and standard deviations – of the descriptor are extracted from the channels shown in Figure 7 in order to model human visual system (HVS). Since text should posses texture that is distinguishable from surrounding to be readable by humans and HTD is based on the HVS model, it has been chosen for feature extraction.

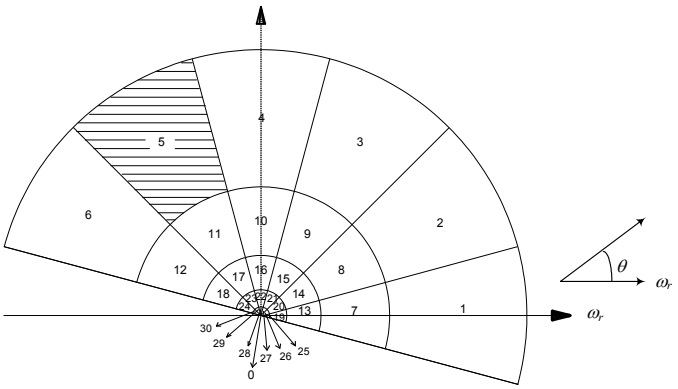



**Figure 7: Homogeneous Texture Descriptor Frequency Partition [25]**

### *2.1.4 Text / Non-Text Classification*

In this section, classification methods considered for overlay-text detection are examined. In the classification process, feature vectors due to one of the previously mentioned methods will be used. Each pattern extracted from 8×8 blocks will be assigned to one of the text or non-text *classes*.

#### *2.1.4.1 Bayesian Minimum Error Rate Classification*

In order to minimize the error rate, Bayes decision rule selects the class with the maximum a posterior probability (2.5).

$$\text{Assign } \omega_i \text{ to } x \text{ if } P(\omega_i \mid x) > P(\omega_j \mid x) \text{ for all } i \neq j \qquad (2.5)$$

Such a decision minimizes the probability of deciding wrong class for the pattern, thus the average error rate is clearly minimized [37]. The form of decision rule can be further arranged by using Bayes' formula:

$$\text{Assign } \omega_i \text{ to } x \text{ if } p(x \mid \omega_i) P(\omega_i) > p(x \mid \omega_j) P(\omega_j) \text{ for all } i \neq j \qquad (2.6)$$

Since $p(x)$ term in the Bayes' formula is just a scaling factor it has been eliminated. Hence, the Bayesian classifier in the form of (2.6) is fully determined by the class-conditional density and the prior probabilities of the categories. For class-conditional probabilities, *multivariate normal distribution* is assumed, while the parameters of normal distribution, mean vectors and covariance matrices, are all estimated from the training set via *Maximum Likelihood* estimation [37], as well as all prior probabilities.

### 2.1.4.2 k-Nearest Neighbor

In *k-Nearest Neighbor* (kNN) method feature vector $x$ is classified according to its $k$ nearest neighbors. The category to be assigned is decided by the majority vote of its neighbors. Although this method is conceptually simple, the naive implementation of the method is computationally quite intensive. Since in such an implementation, the distance of every stored feature vector to tested pattern is calculated, the computational complexity of the resulting method is bounded with $O(dn^2)$, where $O(d)$ is the complexity of distance calculation and $n$ is the size of training set. Nevertheless, there are methods [37] for reducing the computational burden, such as using *partial distance*, *parallel implementation,* or *search tree*. Moreover, at the training phase, some of the training data can be *pruned* by eliminating vectors, which are surrounded by the patterns with the same categories, thus decreasing the number of search elements.

### 2.1.4.3 Learning Vector Quantization

Learning Vector Quantization (LVQ) [26] is a supervised learning and pattern recognition technique in which a set of *codebook* vectors are used to define categories and class boundaries. Unlike general vector quantization methods, LVQ does not approximate density functions of categories; instead the class boundaries are approximated by codebook vectors.

After the initialization of codebook vectors in the learning process firstly, for each training pattern $x$, closest codebook vector $m_c$, is determined and this vector is updated according to the following rules [26]:

$$\begin{aligned} m_c(t+1) &= m_c(t) + \alpha(t)\left[x(t) - m_c(t)\right] \\ &\quad \text{if } x \text{ and } m_c \text{ belong to the same class,} \\ m_c(t+1) &= m_c(t) - \alpha(t)\left[x(t) - m_c(t)\right] \\ &\quad \text{if } x \text{ and } m_c \text{ belong to different class,} \end{aligned} \qquad (2.7)$$

where $\alpha(t)$ is called *learning rate*. In this scheme, if the closest codebook vector classifies training sample correctly, it is moved closer to the pattern, whereas if it does not correctly classify, it is *punished* by moving codebook farther apart from the pattern. Since there is no rule for adapting codebook vectors for unmatched vectors, learning method of LVQ is regarded as *competitive learning*.

The classification of an input pattern is done by assigning the category of the closest codebook vector to the input pattern, as follows [26]

$$\text{Assign } c = \arg\min_i \left\{ \|x - m_i\| \right\} \text{ to } x, \qquad (2.8)$$

where $m_i$ is one of many codebook vectors and $x$ is the input feature vector. Because of this scheme, the codebook vectors represent boundaries piecewise linearly. Moreover, this leads to the partitioning of the feature space into Voronoi cells, as in Figure 8.

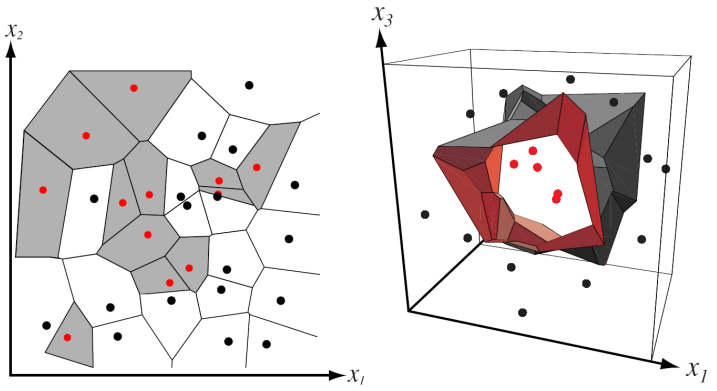



**Figure 8: LVQ codebooks partition feature spaces into Voronoi cells. Simplified representation of Voronoi cells in two- and three-dimensional feature spaces. From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification. Copyright © 2001 by John Wiley & Sons, Inc.**

LVQ can also be thought as a special kind of neural network with two layers [26]. In this perspective, neural network represents codebook vectors and the positions of these vectors are the weights of the neuron connections between input layer and output layer. Although it resembles to *perceptron classifier,* LVQ differs from perceptron on the ability of classifying any set of input vector, not only the linearly separable sets of vectors.

### *2.1.5 Comparative Analysis of Feature Extraction and Classification Methods*

In this section, evaluation results of different feature extraction and classification methods are presented. In the simulations, 240,768 feature vectors extracted from 152 ground-truthed generic video frames are used as training set, whereas the test set consists of 144,144 feature vectors obtained from 91 captured frames. For this analysis, *hit-rate* performance metric is selected which, is characterized as the *correct* assignment percentage of labels to feature vectors of respective classes namely, text and non-text. In Table 1, hit-rate results of the feature extraction – except Haar Wavelet – and classification methods are shown. As it can be observed from the results, DCT-based feature extraction shows consistent and robust results, whereas Edge Histogram has the most inferior performance.

**Table 1: Text and non-text hit-rates of corresponding feature extraction and classification methods. THR: Text Hit Rate, NHR: Non-Text Hit Rate**

| | **Bayesian** | | **kNN** | | **LVQ** | |
|---|---|---|---|---|---|---|
| | THR% | NHR% | THR% | NHR% | THR% | NHR% |
| **DCT** | 77.52 | 89.02 | 73.03 | 85.56 | 76.99 | 83.38 |
| **EH** | 8.31 | 94.63 | 73.88 | 61.47 | 53.14 | 78.47 |
| **HTD** | 87.46 | 51.79 | 73.23 | 84.24 | 83.17 | 63.88 |

Further analyses are performed to select features from Haar Wavelet Transform. The reason of these analyses is to maximize the classification performance while keeping the dimension of the feature vector as small as possible. During the experiments, as a first step, mean, second central moment and third central moment with all sub-bands in the picture are interchangeably used to test their saliency. The results in Table 2 indicate that mean and the second central moment mostly performs better compared to other combinations.

**Table 2: Text and non-text hit rates obtained by using features extracted from sub-bands of Haar Wavelet Transform (HWT). THR: Text Hit Rate, NHR: Non-Text Hit Rate**

| | **Bayesian** | | **kNN** | | **LVQ** | |
|---|---|---|---|---|---|---|
| **Features** | THR% | NHR% | THR% | NHR% | THR% | NHR% |
| $\mu$ | 56.71 | 91.05 | 71.61 | 82.80 | 79.59 | 74.00 |
| $\sigma^2$ | 44.34 | 93.49 | 82.80 | 81.86 | 88.81 | 74.10 |
| $\mu_3^2$ | 31.97 | 96.29 | 77.90 | 84.87 | 88.51 | 77.29 |
| $\mu$ **and** $\mu_3^2$ | 49.41 | 93.41 | 78.13 | 84.44 | 90.10 | 73.68 |
| $\sigma^2$ **and** $\mu_3^2$ | 42.92 | 94.12 | 75.40 | 86.09 | 84.25 | 81.28 |
| $\mu$ **and** $\sigma^2$ | 59.31 | 91.43 | 84.08 | 80.34 | 86.92 | 78.26 |

Secondly, the sub-bands are investigated. In each test case, features are extracted as mean and second central moment from the specified sub-bands. In Table 3 analysis results are tabulated, as well as the feature extraction method described in [2].

**Table 3: Text and non-text hit rates obtained by using feature vectors extracted from different sub-bands of Haar Wavelet Transform (HWT). THR: Text Hit Rate, NHR: Non-Text Hit Rate**

| | Bayesian | | kNN | | LVQ | |
|---|---|---|---|---|---|---|
| **Sub-Bands** | THR% | NHR% | THR% | NHR% | THR% | NHR% |
| **Except LL** | 57.99 | 92.28 | 75.46 | 87.23 | 83.68 | 80.01 |
| **Only HH** | 49.54 | 93.82 | 87.77 | 80.03 | 92.43 | 74.32 |
| **Only LH** | 38.59 | 96.01 | 81.18 | 84.22 | 90.37 | 71.41 |
| **Only HL** | 39.74 | 94.06 | 70.67 | 81.31 | 84.52 | 77.10 |
| **LH and HL** | 49.41 | 93.37 | 76.82 | 84.44 | 80.67 | 83.07 |
| **Doermann[2]** | 42.92 | 94.12 | 75.40 | 86.09 | 83.10 | 81.68 |

### *2.1.6 MRF-Based Text Classification*

In most of the prior work on overlay-text detection, the multi-part nature of the text object is covered by either imposing some heuristics or utilizing neighborhood connected component analysis. Although, in the literature, multi-part-based approach is used for object detection, this approach is not employed in text detection problem, except for the approach in [6]. Moreover, none of the prior work has modeled the objects, as a multi-part object based on a probabilistic framework.

Markov Random Field approaches (MRF) have found broad application areas in computer vision and image processing [27-30]. This approach has been extensively used for low-level image segmentation, image restoration, edge detection and modeling of textures and objects. In [27], Deng and Clausi proposed an unsupervised image segmentation method based on image features by using a simple MRF model. Moreover, in [28], for multi-textured images, Melas and Wilson proposed Double Markov Random Field Models to tackle the

segmentation problem. In object detection applications, a parts-based model is used by Ioffe and Forsyth [29], in order to detect human body, in which boosting is used to combine weak classifiers, corresponding to different body parts. Moreover, in [30], tree structures are used to model objects. Afterwards, the objects are detected in the input images by matching these trained models.

### *2.1.6.1 Preliminaries*

Let $S$ be a regular lattice for a 2D image of size $W \times H$ defined as $S = \{(i, j) \mid 1 \le i \le W, 1 \le j \le H\}$, then any mapping $N : S \to 2^S$ is a neighborhood system, if for all $s, r \in S$, $r \in N_s \Leftrightarrow s \in N_r$ where $s \notin N_s$. Given a lattice $S$ and a neighborhood system $N$, a clique $c$ is any set of lattice sites $c \subset S$ such that $\forall s, r \in c, r \in N_s$. Finally, let $X_s \in \Omega$ be a random field defined on the lattice $S$ with the neighborhood system $N_s$ and assume that $X$ has probability $P(X = x)$ then $X$ is a Markov Random Field, if $P(X = x) > 0 \quad \forall x \in \Omega$ and $P\left(X_s \mid X_{S-\{s\}}\right) = P\left(X_s \mid X_{N_s}\right)$.

### *2.1.6.2 Proposed Approach*

Text detection can be considered as a labeling problem in which the label set consists of two labels, as *"non-text"* and *"text"*. This set of labels is assigned to each block of the image, which form the necessary lattice *S* in our case. Note that the multi-part structure of text causes labels of the blocks (lattice elements) to have some mutual relationship. MRF formulation might provide a solution to incorporate this mutual dependency between labels of the lattice sites. From

Bayesian point of view for the classification, given the observation *x*, the decision of the label *l*, one may simply determine from the conditional probability:

$$p(l \mid x) = \frac{p(x \mid l)\, p(l)}{p(x)} \tag{2.9}$$

Since *p(x)* is same for all it is merely a normalizing factor, it can be eliminated from (2.9) to yield after taking natural logarithm of the following relation:

$$\ln\left(p(l \mid x)\right) = \ln\left(p(x \mid l)\right) + \ln\left(p(l)\right) \tag{2.10}$$

At this point, assuming the posterior density is Gibbsian, each term in (2.10) can be obtained as energies, namely *posterior energy*, *likelihood energy* and *prior energy*, respectively, as:

$$U(l \mid x) = U(x \mid l) + \lambda U(l) \tag{2.11}$$

The minimal solution, $l^* = \arg\min_f U(l \mid x)$, is the solution for this classification problem.

The prior energy can be modeled as a Multi-Level Logistic (MLL) model [31]. With this model, the prior energy *U(l)* can be defined by incorporating clique potentials $V_c(l)$, such as:

$$U(l) = \sum V_c(l), \tag{2.12}$$

with

$$V_c(l) = \begin{cases} +\gamma_c & \text{if all sites on c have the same label} \\ -\gamma_c & \text{otherwise} \end{cases}, \tag{2.13}$$

where $\gamma_c$ is the type-c clique potential. In the MLL model and thus in the MRF model, 8-neighborhood (second order) mode are utilized. The corresponding cliques are illustrated in Figure 9.

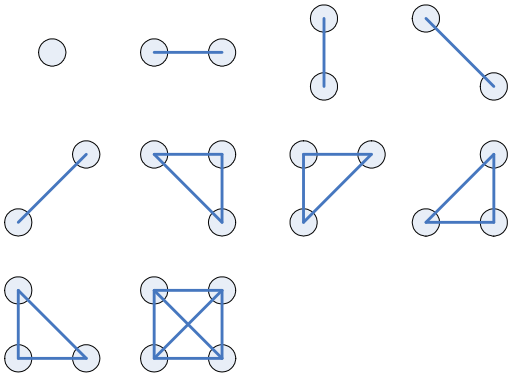


**Figure 9: First-order and second-order mode cliques.**

In the proposed approach, the likelihood energy is derived from a univariate Gaussian distribution. In (2.14), $\mu_l$ and $\sigma_l^2$ are mean and the variance of feature, $x$, given the corresponding label $l$.

$$U(x \mid l) = \frac{(x-\mu_l)^2}{2\sigma_l^2} + \ln\left(\sqrt{2\pi}\sigma_l^2\right) \qquad (2.14)$$

In our realization, the output likelihoods of the Bayesian Classifier for every feature vector are used as observations, $x$, in the corresponding likelihood energy. These feature vectors are extracted by using DCT-based method discussed in Section 0. Iterated Conditional Modes (ICM) algorithm [31], which sequentially minimizes the local energy, is used to obtain the solution for this minimization

problem. It should be noted that minimization results of ICM have significant dependence on the initial estimates of the solution. In the proposed system, for the initialization the minimum error-rate classification results, which are obtained from Bayesian classification, are simply utilized.

### *2.1.6.3 Experimental Results*

For the training process and testing, video frames (352×288), captured from typical TV broadcast is used. The training dataset contains 157 frames, whereas the test set consists of 92 generic frames. In this study, a simple (block-based) precision and recall criteria are utilized. Since the overall system does not aim to detect text boxes, but only the text regions, this performance evaluation scheme is still acceptable. The following precision-recall curve in is obtained by varying the smoothness parameter $\lambda$ in Equation (2.11). In this figure, the blue dots represent MRF results, whereas single red dot shows the precision-recall generated from Bayesian Minimum Error-Rate classification. Moreover, in Figure 12 some typical sample results are shown.

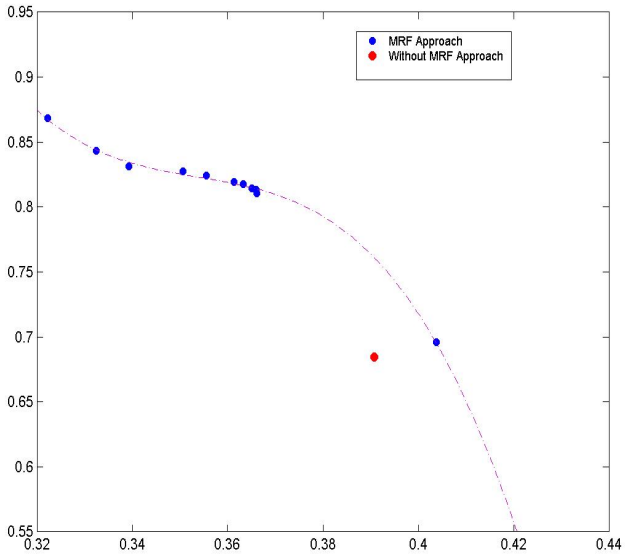



**Figure 10: Precision vs. Recall plot for MRF-based and Bayesian Minimum Error-Rate classifier**


**(a)** **(b)**

**(c)**

**Figure 12: (a) Sample image from test set, (b) initial labels and (c) MRF-based classification result**


**(a)** **(b)**

**(c)**

**Figure 12: (a) Sample image from test set, (b) initial labels and (c) MRF-based classification result**

#### *2.1.6.4 Discussions*

In this part of the thesis, the multi-part structure of the text regions is exploited by utilizing the output likelihoods of Bayesian Classifier in a MLL model as observations. Furthermore, the performance of the proposed method is compared against naïve Bayesian Classifier. Based on the simulation results, the recall performance of the proposed method shows considerable improvements. On the other hand, the precision values do not achieve the same amount of improvement, since the energy model smoothes the boundaries considerably, dilating the text-based regions. Although ICM has proved to be useful for the solution of the minimization problem, its dependence on the initial condition and its deterministic nature forces investigating different minimization methods, including stochastic methods, as a future work. Proposing more sophisticated energy models, which considers different classifier outputs and undesired dilation effect, also remains to be a future research.

### *2.1.7 Bounding-Box Extraction*

At the end of text/no-text classification, the results are still not appropriate for being processed by character recognition module. For this reason, the first level classification results are transformed into separate *word* regions by using bounding-box extraction technique. During this process, connected-component labeling, edge detection and projection profiles are used.

#### *2.1.7.1 Connected-Component Labeling*

Connected-component labeling (CCL) [32] algorithm seeks to assign a unique label to each connected object. Typically, CCL is applied to a binary image. In our case, the classification result produces this binary map in which each block is

designated as text or non-text. Although there are different algorithms that solve CCL problem, following sequential algorithm is preferred:

***Algorithm 1: (Sequential Connected-Component Labeling)*** **[32]**

*1* Scan image left to right, top to bottom
*2* If the pixel is 1, then
*3* If only one of its upper and left neighbors has a label, then copy the label
*4* If both have the same label, then copy the label
*5* If both have different labels, then copy the upper's label and enter the labels in the equivalence table as equivalent labels.
*6* Otherwise assign a new label and enter this label to equivalence table
*7* If there are more pixels to consider, then go to step 2
*8* Find the lowest label for each equivalent set in the equivalence table.
*9* Scan the image. Replace each label by the lowest label in its equivalent set.

### *2.1.7.2 Sobel Operator*

Sobel operator [32] is simply the magnitude of the gradient computed at each point of the image by:

$$M = \sqrt{s_x^2 + s_y^2}, \qquad (2.15)$$

where each of the gradients $s_x^2$ and $s_y^2$ are computed by:

$$s_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} * A \qquad s_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} * A, \qquad (2.16)$$

where $A$ is a 3×3 neighborhood of the current pixel in source image.

### *2.1.7.3 Method for Bounding-Box Extraction*

Firstly, text-labeled regions are grouped by CCL. At this level, the components which are smaller than some certain threshold value are filtered. Furthermore, after fitting bounding-rectangles on every component, the ones that have fill-factor smaller than a certain threshold value are also removed. In the next step, edge maps of every connected component are extracted by using Sobel operator [32]. In every edge map, projection-profiles are computed in *x*- and *y*-directions. By using these profiles, text lines and eventually words are segmented. In Algorithm 2 the details of the proposed method are provided.

***Algorithm 2: (Bounding-Box Extraction)***

*1* Scan each connected-component in the list
*2* Compute fill-factor and area for the component
*3* If fill-factor is smaller than a threshold and area is smaller than a threshold go to step 2
*4* Take *x*-gradient Sobel of component
*5* Compute projection profile in vertical direction
*6* If there is disjunction divide and add every dividend to line set.
*7* Else add component to line set
*8* If there are more components continue from step 2
*9* Else for every element in line set
*10* Take Sobel
*11* Compute projection profile in horizontal direction
*12* If there is disjunction divide and add every dividend to word set
*13* If there are more components continue from step 8
*14* End

In Figure 13, a typical example for bounding-box extraction with its intermediate steps can be observed.

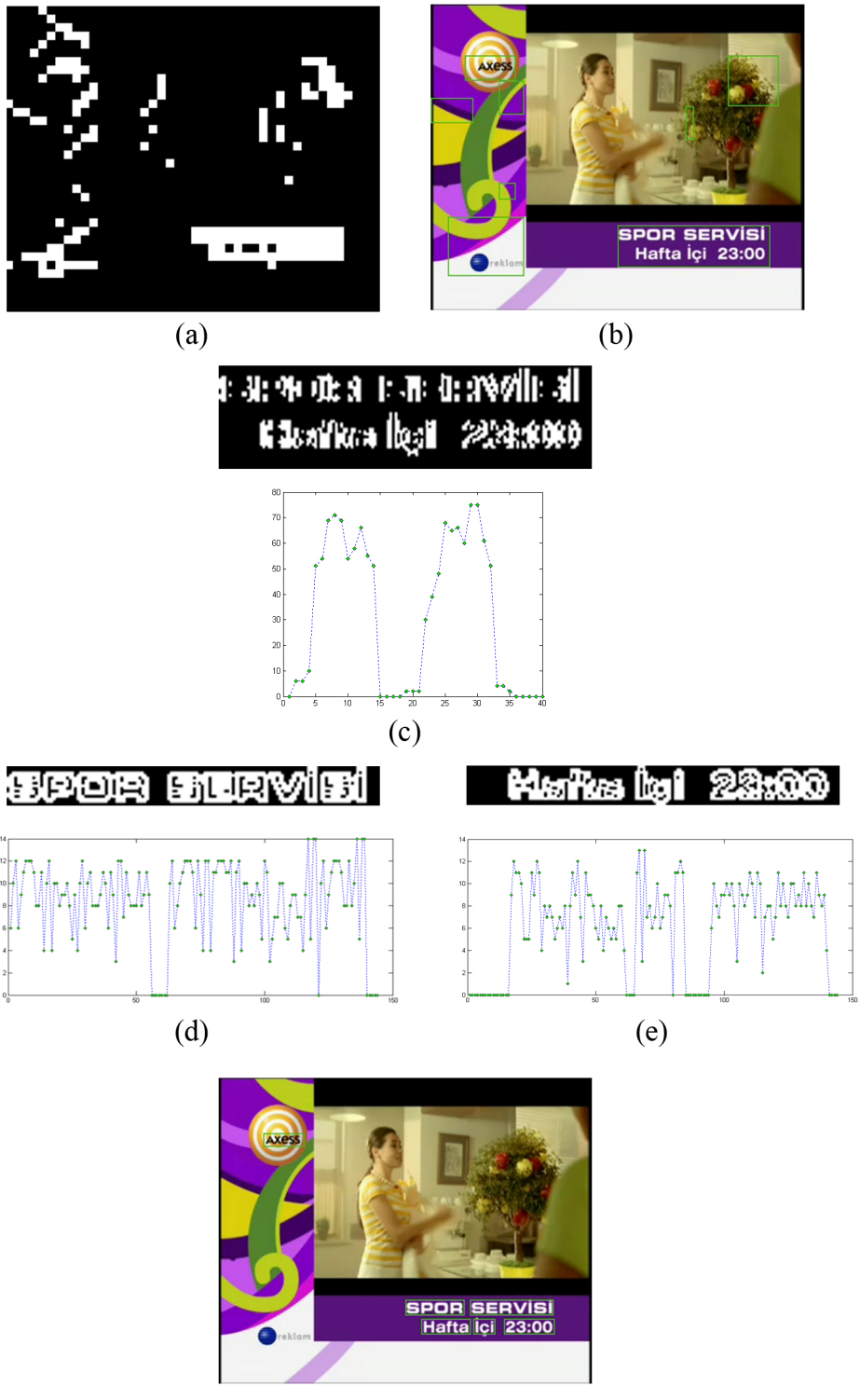



**Figure 13: (a) Classification Results; (b) Connected-Components (green rectangles) after pruning; (c) Single component's *x*-gradient Sobel Edge Map and its horizontal projection profile; (d) and (e) Sobel Edge Maps and vertical projection profiles of lines extracted from (c); (f) Final results depicted by green bounding rectangles.**

### *2.1.8 Word Bounding-Box- Localization Evaluation*

In this evaluation scheme, there are three performance criteria, as *Hit Rate*, *Error Rate*, and *False-Alarm Rate*. In Figure 14, the regions that are essential for the performance analysis of the algorithms are depicted; in which the *Ground Truth Rectangle* term corresponds to the actual text-box and *Detected Rectangle* represents the result for the evaluated algorithm.

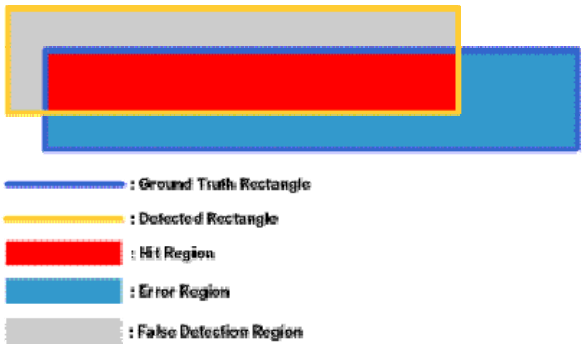



**Figure 14: Bounding Box Evaluation.**

*1. Hit Rate:*

*Hit Rate* (HR) is the ratio of the area of *Hit Region* in Figure 14 to the corresponding *Ground Truth Rectangle*'s area. In a more formal way, let *d* be the detected region, *g* be the ground truth region and *o* be the *Hit Region* (overlapped region) corresponding to a matched ground truth. Then HR is calculated as

$$HR = \frac{\sum_{g \in T_g} \frac{A(o)}{A(g)}}{N} \qquad (2.17)$$

where $T_g$ is the set of all ground truth rectangles, $N$ is the number of total ground truths in a frame and $A(\bullet)$ is the area operator.

2. *Error Rate:*

*Error Rate* (ER) is the ratio of the area of *Error Region* to the corresponding *Ground Truth Rectangle*'s area that is the proportion of the region that is not detected by the method to the total area of the ground truth regions. Similar to HR, in a more formal way, with the same initializations; ER is also given as

$$ER = \frac{\sum_{g \in T_g} \frac{A(g - o)}{A(g)}}{N} \qquad (2.18)$$

3. *False Alarm Rate:*

Finally, *False Alarm Rate* is the ratio of the *False Detection Region*'s area to the whole client area that is the total area of the frame; in other words, the proportion of the region that is erroneously stated as a text region in the whole frame. Again with the same initializations the formal expression of *False Alarm Rate* is equal to

$$FAR = \frac{\sum_{d \in T_d} A(d - o)}{A(I)} \qquad (2.19)$$

For the evaluation of the overall system, a challenging dataset is composed from 21 Turkish television channels. 34 hours of data is captured in H.264 with 500 kbps rate and the resolution of the videos is selected as 352×288 from these channels. Moreover, out of this pool, 962 frames are ground-truthed generating a total of 11872 text regions. The evaluated system is made up of aforementioned DCT-based feature extraction and Bayesian classifier coupled with the Bounding-Box Extraction method described in 2.1.7.3. Finally, Hit-Rate is calculated as

67.51% and False-Alarm as 36.22% whereas Error-Rate is calculated to be 32.49%.

## *2.2 Overlay-Text Recognition*

At the end of the localization phase, the regions belonging to individual words are encapsulated by bounding-boxes. By the subsequent recognition module, the image content under each bounding-box is processed to identify the underlying text information.

At this point, it should be emphasized that successful Video OCR applications require more than off-the-shelf OCR modules. High performance Video OCR solutions usually require tuning and various customizations to the OCR. There are many reasons for this conviction. Firstly, most of the off-the-shelf OCR packages require high-resolution data, which is not the case in Video OCR problem. Moreover, most of the OCR packages does not handle non-uniform background problem, since printed or handwritten documents has fairly uniform background.

The overall recognition system used in this work is derived from *NIST Form-based Handprint Recognition System* [32]. The components of the recognition system are depicted in Figure 15, whereas their details are given in the following sub-sections.

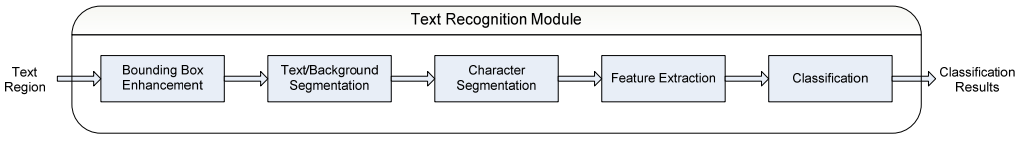



**Figure 15: Block Diagram of the Recognition Module.**

### 2.2.1 Enhancement

When the localization process ends, the resulting text regions can have resolution below the permissible recognition range. With an enhancement stage, low-resolution can be handled. Although super-resolution (SR) techniques can be employed for the resolution enhancement, simple interpolation techniques are utilized in this thesis. This selection is due to a number of reasons, such as the simplicity of interpolation techniques compared to SR techniques and the requirement of multiple-frames and/or a form of aliasing by SR methods to be effective, which may or may not be present at hand.

#### 2.2.1.1 Bilinear Interpolation

During bilinear interpolation, the value of the output pixel is estimated by using its four nearest neighbors; Figure 16 illustrates the process. In this approach, linear interpolation is first performed in one direction and then that result is interpolated linearly in the other direction (2.20).

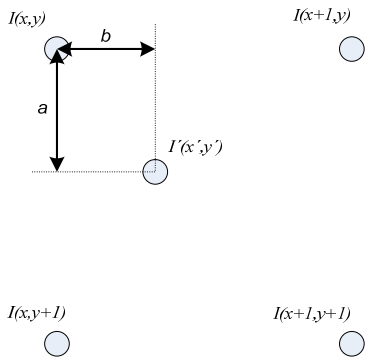



**Figure 16: Bilinear Interpolation**

$$I'(x',y') = (1-a)\left[(1-b)I(x,y) + bI(x,y+1)\right] + a\left[(1-b)I(x+1,y) + bI(x+1,y+1)\right] \quad (2.20)$$

The step-like block effect of nearest neighbor interpolation is solved with this method. However, some blurring is introduced with this interpolation technique.

### *2.2.1.2 Bicubic Interpolation*

Interpolation techniques can also be expressed compactly by introducing interpolation kernels:

$$I'(x',y') = \sum_{m=-1}^{2}\sum_{l=-1}^{2} I(x+l,y+m)\,u\left(x'-(x+l)\right)u\left(y'-(y+m)\right); \quad (2.21)$$

where *u(s)* is an interpolation kernel.

Bicubic interpolation, as introduced in [34], improves estimation of the target pixel value, not only by using 16 neighboring pixels in the source image, but also approximating *sinc* reconstruction function with a better model.

$$u(s) = \begin{cases} \frac{3}{2}|s|^3 - \frac{5}{2}|s|^2 + 1 & 0 < |s| < 1 \\ -\frac{1}{2}|s|^3 + \frac{5}{2}|s|^2 - 4|s| + 2 & 1 < |s| < 2 \end{cases} \quad (2.22)$$

Kernel function defined in (2.22) is a continuous piecewise polynomial and has a continuous first derivate, which makes it an attractive candidate. Figure 17 illustrates the improvement, as less blurring and preservation of fine details, accomplished by Bicubic interpolation over bilinear interpolation.


**Figure 17: Resolution enhancement results by a factor of 6 (a, c) and 8 (b, d) by using bilinear interpolation – c and d – and bicubic interpolation – a and b.**

### *2.2.2 Text / Background Segmentation*

Unlike document recognition, in Video OCR noise and complex background are commonly observed defects. Thus, for an efficient recognition, segmentation of text characters from noise and background is vital.

**Table 4: Recognition rates**

| Method | Recognition Rate |
|---|---|
| Entropy_Kapur | 0.756 |
| Local_Bernsen | 0.688 |
| Cluster_Lloyd | 0.679 |
| Entropy_Li | 0.674 |
| Entropy_Shanbag | 0.674 |
| Entropy_Sahoo | 0.665 |
| Cluster_Yanni | 0.662 |
| Cluster_Otsu | 0.659 |
| Cluster_Ridler | 0.655 |
| Cluster_Jawahar | 0.655 |
| Shape_Ramesh | 0.613 |

In the previous preliminary analysis (Table 4), character recognition rates of word regions, which are binarized by various thresholding [35] techniques, are obtained. During this comparison, 12 word regions with a total of 76 characters are processed.

#### *2.2.2.1 Multi-Hypotheses Approach for Text vs. Background Segmentation via k-Means Algorithm*

Although most of the previous video text recognition methods proposed in the literature strive to improve binarization results, finding the optimal binarization solution is difficult to achieve, due to the reasons, such as complex background, noise and several grayscale distribution modes around the word region. In order to overcome this difficulty, instead of segmenting word region into two levels, multi-level segmentation is proposed. For each segmented level, a binary hypothesis is constructed by assuming corresponding level as text and all other levels as background. Finally, all of these hypotheses are fed to the following stages. In the related literature, Odobez et al. has proposed Expectation Maximization and Gibssian EM-based methods in [36]. However, in this work for multi-hypotheses generation *k-Means* is employed.

#### *2.2.2.2 k-Means Clustering*

k-Means is one of the simplest clustering methods used widely, which minimizes the following objective function:

$$J = \sum_{i=1}^{k} \sum_{x \in D_i} \left\| x - \mu_i \right\|^2, \qquad \text{where } \mu_i = \frac{1}{n_i} \sum_{x \in D_i} x. \tag{2.23}$$

The steps of the algorithm are as follows:

***Algorithm 3: (k-Means Clustering)*** **[37]**

*1* Arbitrarily initialize *k* cluster centers,

*2* Classify each sample according to the nearest cluster center,

*3* Re-compute each cluster center(mean),

*4* Stop if cluster centers do not change otherwise go to step 2.

For the sake of completeness following remark is necessary. The resulting clustering depends on the initial values of the clusters, and the suboptimal partition can be obtained frequently.

**(a) (b) (d)**

**Figure 18: (a) Source image; (b) Thresholding result by Kapur; (c) k-Means based segmentation result with 3 clusters.**

**(a) (b) (c)**

**Figure 19: (a) Source image; (b) Thresholding result by Kapur; (c) k-Means based segmentation result with 3 clusters. Characters *p* and *t* emerge as touching characters**

In Figure 18 and Figure 19, sample segmentation results are depicted. In Figure 18, brightness change in the background cause loss of characters at simple thresholding. However, by employing k-Means based method instead, none of the characters are lost due to segmentation. Furthermore, coinciding characters (Figure 19b); after thresholding can be corrected by k-Means (Figure 19c).

### *2.2.2.3 Evaluation*

The aforementioned k-Means based Text/Background Segmentation is compared with the best performing conventional segmentation technique from Table 4 in order to demonstrate the multi-hypothesis approach's success. Assessment is carried out on individual words and evaluation is based on the recognition performance on these words. Test set, extracted from generic TV captures, contains 8085 individual word regions and a total of 44288 characters.

In the assessment, initially following metrics are considered; *Perfect Match:* Percentage of whole word recognition; *Correct Characters:* Percentage of correctly recognized characters at the correct position in the word. Furthermore, with the purpose of analyzing system more accurately three more metrics; *Precision, Recall* and *Accuracy* are calculated by using *Levenshtein Distance.* Precision is calculated as the ratio of matching characters and the total number of characters recognized for the given word, whereas the Recall is the ratio of matching characters and the total number of characters in the ground-truth. Moreover, Accuracy (*f*) is calculated from these two metrics as follows;

$$f = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (2.24)$$

The obtained evaluation values are shown in Table 5. As it can be observed from the results, multi-hypothesis approach improved overall results by approximately 15%.

**Table 5: Individual word recognition assessment for Text/Background Segmentation.**

| | Kapur | k-Means |
|---|---|---|
| **Perfect Match** | 37.7% | 48.1% |
| **Correct Characters** | 57.6% | 68.6% |
| **Precision** | 71.2% | 81.1% |
| **Recall** | 65.9% | 76.3% |
| **Accuracy** | 68.4% | 78.6% |

### *2.2.3 Character Segmentation*

There have been continuous efforts for developing and improving character segmentation methods for character classification. However, character segmentation still remains as a source of errors, especially for Video OCR. The inherent problems of Video OCR, such as low-resolution and arbitrary size of text in video, produce the problem of coinciding characters.

In the literature, many researchers have proposed methods to solve the segmentation problem. In [41], Casey and Lecolinet present an analysis of the previous methods. Rocha et al. in [38] propose character segmentation technique in which saddle features between characters, which are identified in grayscale, are used for segmentation. Another approach proposed by several researchers is to segment characters by a recognition-based technique; such as the published works of Keeler [40] and Matan [39].

In this thesis, the method used at the character segmentation stage is directly obtained from [33] with minor modifications. In this method, as a first step, connect-components of the binarized word region image are extracted. These components most of the time represent single complete characters. Based on connected-components, further improvements are achieved in order to handle

noise, fragmented and/or touching characters by using a simple adaptive writing model. Finally, all extracted character regions are normalized in order to prepare characters for feature extraction stage.


(a) (b)

(c)

**Figure 20 (a) Source grayscale image; (b) k-Means result; (c) Resulting characters after character segmentation and normalization**

### *2.2.4 Feature Extraction*

Feature extraction is carried on the binary character images that are obtained at the previous step. The Karhunen Loève (KL) transform is adopted for feature extraction in order to reduce dimensionality and noise. From the set of training characters, a covariance matrix and its corresponding eigenvectors are computed. Each eigenvector and the character to be recognized is taken as an inner product to obtain a coefficient of the feature vector and by combining these coefficients 128 length feature vector is obtained.

### *2.2.5 Classification*

In order to classify the feature vectors for the characters, Multi Layer Perceptron [37] (MLP) based system is adopted. MLP used in the system consists of input

and an output layer, as well as a single hidden layer. Training of the neural network is carried out by *conjugate gradient descent* [37], which employs a series of line searches in weight and parameter space. Moreover, *Boltzmann pruning* [37] is used to constrain the dimension of the weight space. In the training of the neural network, *sinusoidal* activation function is selected; furthermore, the number of the hidden units in the hidden layer is chosen as 384. Since feature vectors are 128-dimensional, the input layer consists of 128 units. Finally, the output layer contains 41 neurons. This reduced number of output neurons is result of output class unification. Some of the output classes such as uppercase *c* "C" and lowercase *c* "c", are unified in both training and classification stages in order to decrease the average error of the network. For the training process 36,093 patterns are used and average percentage of the correctly identified categories at the end of training is computed as 98.66%, whereas the worst classified category has performance of 94%.

### *2.2.6 Post-Processing*

Although neural network has outstanding ability to classify single characters, due to text localization and capture errors recognition still brings significant problems. Even if character recognition accuracy rate reaches as high as 99%, only a 95% word recognition accuracy rate will be reached, assuming an average of five character words. As a consequence numerous research efforts have been addressed to post-recognition spelling correction. For a thorough analysis of the literature on this subject, the reader is referred to [42].

In our approach, a dictionary-based ranking mechanism is proposed. Generally, dictionary based methods directly use the suggestions from the dictionary. However, a more convenient way would be incorporating this suggestion with the confidence information present in the neural network outputs. In this way, the unnecessary suggestions can be eliminated. Given a word to be corrected the

system first obtains candidates from the dictionary by using *Levenshtein Distance* [43] string matching. Afterwards, each candidate's distance to the recognized character is calculated in terms of *deletion, replace* and *insert.* Hence, according to the cost function in (2.25), for each candidate a cost is calculated. For the calculation of this cost function, the confidence values obtained from artificial neural network and some observations are used. Deletion cost of a character is taken as the confidence value of the character. Replacement cost is obtained from the difference between the confidences of these characters. Insertion cost, however, is taken as the confidence value of the character at the position of insertion, if there is a character.

$$C = \sum_{d \in D} \lambda_d c_d + \sum_{r \in R} \lambda_r c_r + \sum_{i \in I} \lambda_i c_i + \sum_{h \in H} \lambda_h c_h \qquad (2.25)$$

In Figure 21, the aforementioned system is depicted.

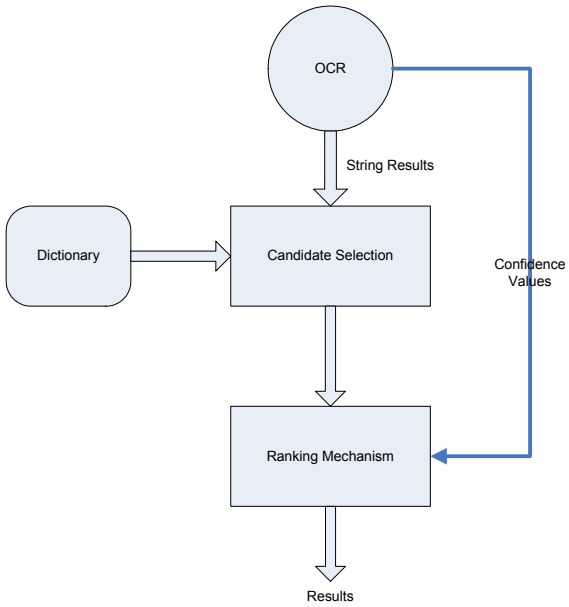



**Figure 21: Post-Correction Module.**

During the experiments, first, without using ranking mechanism, ranks of corrected words in the suggestion set are recorded. Secondly, with ranking mechanism activated ranks of the corrected words in the suggestion set are also compared.

**Table 6: Ranking mechanism results.**

| | **Correction Results without Ranking Mechanism** | | **Correction Results with Ranking Mechanism** | |
|---|---|---|---|---|
| **Rank** | # | % | # | % |
| **1** | 803 | 55 | 1008 | 69 |
| **2** | 202 | 14 | 176 | 12 |
| **3** | 102 | 7.0 | 84 | 5.8 |
| **4** | 88 | 6.0 | 63 | 4.3 |
| **5** | 54 | 3.7 | 38 | 2.6 |
| **6** | 48 | 3.2 | 27 | 1.9 |
| **7** | 41 | 2.8 | 24 | 1.6 |
| **8** | 51 | 3.4 | 18 | 1.2 |
| **9** | 41 | 2.8 | 15 | 1.0 |
| **10** | 32 | 2.1 | 9 | 0.6 |
| **Total** | 1462 | 100 | 1462 | 100 |

## *2.3 Combined Evaluation*

Ultimately, entire Video OCR system with all of the aforementioned building blocks is tested against a considerable amount of video data, captured as discussed in Section 2.2.2.3. However, in this case, somewhat slightly trimmed subset of our previously mentioned data is used; regions containing numerical characters are removed from recognition evaluation. Moreover, both text localization and recognition performances are monitored in this analysis based on the discussions in Section 2.1.8 and Section 2.2.2.3. However, as an important note, for the overall recognition performance only the detected regions that

overlap sufficiently with the corresponding ground-truth are considered. In Table 7 detailed results of evaluation is provided. Additionally, the number of perfect matching words is determined to be 2706 out of 6973 words.

**Table 7: Overall performance results obtained.**

| | **Localization** | | **Recognition** |
|---|---|---|---|
| **Hit-Rate** | 67.51% | **Precision** | 70.26% |
| **False-Alarm** | 36.22% | **Recall** | 77.94% |
| **Error-Rate** | 32.49% | **Accuracy** | 73.90% |

# CHAPTER 3

# SCENE-TEXT DETECTION AND RECOGNITION

There have been numerous research efforts that address text detection/extraction throughout the past years. Although most of these attempts have been devoted for the extraction problem of *artificial*, i.e. *superimposed (overlay)* text, there has been significant focus on the *scene* text extraction in the recent years as well. One of the main reasons of this recent development is the fact that the scene-text extraction has a wide range of application areas; for instance, traffic sign recognition [75], foreign sign translation [53],[54],[57],[61],[62],[72], license plate extraction [59],[60] and some other applications [64],[65].

A sign conveys information about destination, direction, warning, name, advertisement; e.g. a traffic sign, a billboard. Since the depicted information most of the time is an important, relevant and crucial information about the scene, extraction and recognition of this information would be indispensable for various applications, such as sign translation from East Asian languages to English [53], [54], [57], [61], [62]. Since the characters from these languages are difficult to understand and directly copy, these assistance systems are proved to be quite useful. Another application area can be automatic systems for visually impaired people [46], [50], [51], [63] with the help of text-to-speech technology. Moreover, these systems can be arranged to work on mobile devices, such as PDAs and/or cell-phones [54], [57], [73], [74].

Although not a recent encountered problem, the increased demand for security and traffic control allowed License Plate Recognition systems [59], [60] to be an important application area of scene text extraction. Parking area systems, highway/bridge toll systems and many more systems have been put into practical use. However, this problem is slightly a simpler version of the general scene text problem, since the text information is located on a quite differentiable background.

Another interesting application area would be visual text input extraction and recognition. By a camera-based system text information can be extracted from presentations or posters [65], white boards [64] for digitization or indexing. One advantage amongst many would be the fact that those systems would not force users to wear or use any extra apparatus.

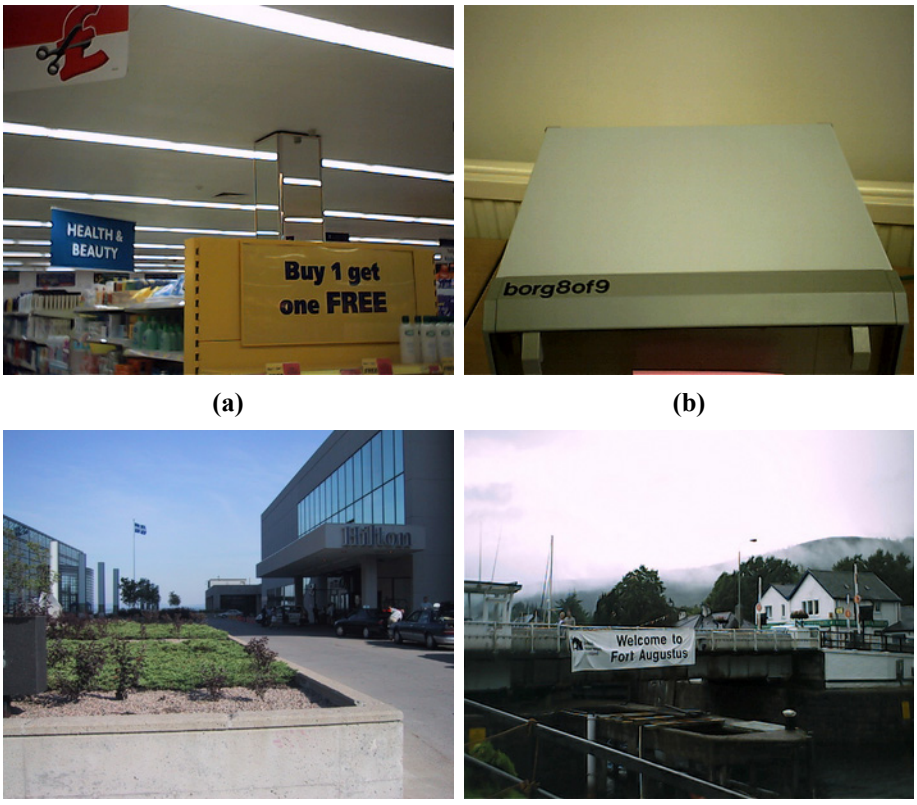
## *3.1 Challenges*

Even though advanced state-of-the-art Optical Character Recognition (OCR) Systems are matured over decades and started to produce good results, there are major difficulties faced by scene text detection and recognition systems that are not handled in traditional OCR problem and overlay text detection and recognition systems. These challenges can be categorized as, Perspective Projection Distortion, Complex Scene, Non-Planar Text Surfaces and Capture Issue(s). These problems are examined in the subsequent sections.

### *3.1.1 Perspective Projection Distortion*

In traditional OCR applications, some of the distortions – rotation, translation, slant and skew – are emerged and handled to some degree; however, in scene-text

detection and recognition distortions cause significant problems to a greater extent. Since scene-text originates in a 3D space, text regions can be distorted by perspective projection, if text plane is not parallel to the image plane. Moreover, in the presence of such distortion, assumption of parallelism is no longer valid, parts of text look shrinked and/or enlarged depending on its distance to the image plane; furthermore, large skew and/or slant distortions might also occur (Figure 22-a and Figure 22-b). In scene analysis, aforementioned distortion is a typically encountered situation for most of the captured imageries, considering that usually text regions are not the focus of intention in the scene (Figure 22-c and Figure 22-d). Even if the text regions are intended as main subjects of the scene, there will be slight distortions due to misalignment of image plane to the text plane.




**(a)** **(b)**

**(c)** **(d)**

**Figure 22: (a) – (d) Example images containing text from ICDAR 2003 Text Locating Competition.**

### 3.1.2 Non-Planar Text Surfaces

Unlike traditional document analysis with flatbed scanner systems or overlay-text detection and/or recognition; scene-text can be observed on any surface (Figure 23); thus, affecting detection and recognition severely in case of non-planar surfaces. A slightly relating problem arises in traditional OCR problem, during the scan of a book with flatbed scanner. In such a case, page curls can be formed at the binding of the book; thus, producing a cruddy image. Although domain specific approximation to text surfaces can be used to solve such a specific problem, a generic solution is not present for camera-based text detection/recognition.


**Figure 23: Scene-text regions on non-planar surfaces.**

### 3.1.3 Complex Scene

In scene-text detection, opposed to document analysis, most of the captured image area does not contain text, and again on contrary to overlay text detection (or document analysis) text regions, mostly, are not positioned on uniform backgrounds. Over and above of all these conditions, the simple fact that scenes can contain almost anything turns/transforms the problem of detection and/or

segmentation a major bottleneck (Figure 22, Figure 23 and Figure 24). In other words, the overlay text is usually imposed in front of a background yielding a higher contrast, whereas the background of typical scene text content is quite arbitrary and makes recognition difficult.

**Figure 24: Complex Scene example**

### *3.1.4 Capture Issues*

Most of the OCR systems work on high resolution data, between 150 dpi and 400 dpi, which is acquired in a controlled environment. However, camera captured image resolutions usually vary below 150 dpi; additionally, uneven lighting conditions, while physical handicaps, such as shadows and reflections, are common situations. Furthermore, uneven focus and sensor noise can cause significant problems compared to flatbed scanners.


**Figure 25: Example images depicting various capture issues.**

All of these problems and conditions also coupled with other *minor* problems, such as occlusion, text orientation and etc. that increase the challenge to construct a scene-text detection/recognition system. However, despite all of these challenges, considering the aforementioned applications, benefits of such a system are obvious.

## *3.2 Related Work Focusing on Scene-text Detection*

As one can deduce from the previous discussion, there is wide range of application areas of scene text detection problem. Whether it is intended to recognize text regions for translation or for the aid of blind, a vital step for the success of the whole system is text extraction (detection) step. Since the scene-text is randomly positioned and/or geometric distortions are present in the acquired image, extraction is more challenging compared to detection of overlay-text or text on structured documents. Although there are efforts that combine both text extraction and recognition steps, in the following sub-section, only scene-text localization/extraction methods are surveyed. According to [66], the algorithms presented in the literature can be roughly classified as *gradient/edge-based, color-based and texture-based methods*. In this survey, a similar classification is adopted; however, it should be noted that most of the work in the literature has tendency of using combinations of these approaches.

### *3.2.1 Gradient/Edge-Based Methods*

The methods in this category generally assume that text characters preserve strong edges against the background regions; thus, most of the proposed algorithms use simple edge or gradient analysis, as a starting point.

Shen et al. in [44] used a graphical model for pruning false alarms obtained from text extraction results. This graphical model is constructed by candidate text regions (features as they referred), which are used to assemble node set of the graphical model. After the extraction of edges, they are combined to form sticks and boxes, which constitute first level features in their work. After the grouping of the first level features, a lot of false alarms are produced. Graphical model constructed prunes false alarms by assuming that correct regions have a regular structure and false alarms have a cluttered distribution. The connections of nodes are achieved within a distance similar to normalize cuts segmentation rather than typical 8-neighborhood. Since the first level features and candidate regions form a sparse graph, inference is calculated by Belief Propagation [28]. In their approach affinity functions are used to measure compatibilities of node elements and the compatibility function is determined by trial and error. Although the results are presented on different images, a thorough analysis of the performance is missing.

In [46], Chen et al. use several weak classifiers, as input to an AdaBoost machine learning algorithm and 4 cascade classifiers, which consist of 79 features, are trained. Moreover in this work, an adaptive binarization method is introduced before the application of recognition phase. First set of features, which form weak classifiers, are based on block mean and standard variation of intensity and modulus of x and y derivative filters. For more complex features, they use histogram of intensity, gradient direction and intensity gradient. Moreover, final set of features are based on edge detection by intensity thresholding followed by edge linking. In summary, (i) 40 first class features including 4 intensity mean features, 12 intensity standard deviation features, 24 derivative features, (ii) 24 second class features including 14 histograms features, and (iii) 25 third class features based on edge linking. In spite of good results reported in this paper, there are also risks of erroneously detecting window frames and building parts as text regions due to the block patterns used as part of the feature extraction. In order to reduce these kinds of errors, the proposed system relies heavily on the OCR package integrated.

In a different approach [50], as a first step, text area with small characters is detected and then it is zoomed into these areas to obtain higher resolution images for character recognition. In the proposed algorithm, four different character extraction methods, based on connected components and as an application system for visually impaired, are presented. From their studies, it has been concluded that a single method can not be enough to extract all text regions in different sizes. For small characters, a modified top-hat processing is applied. This method encounters problems when dealing with low contrast text regions. After the difference image is obtained, a connected component algorithm is used to extract text regions; further improvement is obtained by taking only horizontal areas. On the other hand, for larger characters, three further methods are proposed. In the first one, character extraction is started from Sobel edge maps of RGB channels separately. Single output image is obtained by taking maximum of three edge values on each channel and the output image is binarized. Finally, connected components are extracted. The second method only differs from the first method by reversing the binarized image before connected components are extracted. In the final method, color-based clustering is employed. For every channel, a binary map is obtained by Otsu's thresholding method. Connected components are extracted and selected on each channel separately from these binary images. In all cases, final text regions are extracted by employing thresholds to the properties, such as aspect ratio and area size. Although the aforementioned methods consider different cases, edge-based method outperformed other methods on the ICDAR 2003 dataset. Moreover, as concluded in the paper, the results are still not enough for practical use. Additionally, proposed methods detect and extract scene-text character-by-character which would need a post-processing step to fuse these results into meaningful word results; thus, increasing the chance of overall recognition failure.

### *3.2.2 Color and Texture Based Methods*

Color-based methods are developed under the assumption that any of text should have a distinguishable color or brightness, noticeable with respect to the background. Additionally, texture-based methods view text, as another texture that has certain nature that distinguishes text from background. This assumption can be supported by the fact that humans might locate text of foreign languages without the knowledge of that language, mostly due to its texture.

Jiang et al. [47] decompose the input image is first decomposed into connected-components (CC) by color clustering, which by the way has color tolerance to handle noise and illumination variation on the scene text. These CCs are verified as text regions by a two-stage classifier, a cascade classifier and a Support Vector Machine (SVM). In the two-stage classification process, a total of 15 features are used. These features are divided into 5 groups: geometric, shape regularity, edge, stroke and spatial coherence features. The first stage classifier is a simple feature thresholding method with one upper and one lower threshold. The results are quite successful in a test set of 500 images. Unfortunately, as reported in the paper, the proposed method is not capable of locating text on reflecting surfaces.

In [48], unsupervised clustering based method is used to decompose color image into three-channel (Red, Green, and Blue) images. After this step, a feature vector calculated from standard deviation of 8x8 blocks of LH, HL, HH bands of the first level 2D wavelet transformations performed on the decomposed images. Next, K-means algorithm applied on the normalized 3D feature space in order to obtain 3 clusters, namely background, text and complex background. Although experimental results do not show much improvement on recall rates, introduction of both three color channel processing and complex background concept, reduce the incorrect classification.

In a different approach [54], detection of text regions on signs in various natural scenes is aimed. The approach efficiently embeds multi-resolution adaptive search in a hierarchical framework. Moreover, an intensity-based OCR method is introduced, in which Gabor transform for local features and LDA for selection and classification is used. The detection method used in this system is again inherited from [52], whereas in this study a two cluster background/foreground Gaussian Mixture Model (GMM) [100]is utilized. Additionally, for the layout analysis, a Hough transform is incorporated. For the character recognition step, an intensity and Gabor Jets-based [101] feature extraction method is preferred. Before extraction of features, text normalization is used for characters in order to decrease discrepancies of intensities on character regions caused by different lighting. After this step, an LDA is used to select features and finally, k-NN is used for character classification. Moreover in [55], Yang et al. improve their previous studies [52], [53] and [54] on text detection with the introduction of affine transformation in order to increase text recognition accuracy.

Yet another color clustering-based approach is proposed by Wang et al. in [58]. As a first step, color feature vectors of text regions are clustered into a number of color classes by coarse and fine with the application of fuzzy k-Means algorithm. Then, different results are constructed according to these clusters. Finally, the characters are extracted from the images by using the information of segmentation and recognition. In the coarse classification step, 2D histogram of color feature vectors presented by [68] are segmented by applying topographical features, namely *peak, pit and saddle* [69]. After fine classification, slices corresponding to color classes are computed. The classification of these slices, as characters or background, is achieved by the combination of two criterions, uniformity of sizes/spaces of characters and the generalized recognition confidence. However, proposed method experience difficulties when colors of characters are close to background color and when there are light variances.

Ezaki et al. has further extended a previous work [50] to increase small text character localization performance by using Fisher's Discriminant (FD) method [51]. The image is divided into non-overlapping 32x32 blocks and FD Rate (FDR) is obtained from histogram of each block. With the observation of, histogram has two peaks for text occupied tiles, which yields high value of FDR, whereas for non-text regions only one peak is present. For complex regions, a higher FDR than that of non-text regions is obtained, whereas still a lower value than text regions is resulted. In the proposed method for low FDR valued blocks, a global thresholding and for high FDR blocks a local thresholding is applied. After binarizing all three channels, connected components are extracted independently. In all cases, final text regions are extracted by employing thresholds to the properties, such as aspect ratio and area size as in the previous work. The proposed method fused with the previously reported morphology-based method [50] reaches better recall rates, however overall accuracy of the combined system is still low for practical applications.

As it has been pointed out, some of the approaches in the literature employ both color- and edge-based methods in conjunction. In [52], an adaptive method for detecting text regions from natural scenes is presented. The proposed method first starts with the extraction of initial text candidates from image/video. After this step, an adaptive color modeling and searching algorithm is utilized around the candidate regions to discriminate text/non-text regions. Initial candidate text regions are detected by using a multi-resolution edge filtering-based method. After this step, edges are clustered according to properties, such as aspect ratio, existence of pairs of rising and falling edges, etc. In the next layer, first a search region and an extension region is defined by incorporating layout direction and the candidate regions. After this step, in order to utilize the assumption of color or intensity of text does not change significantly; a Gaussian mixture color model is incorporated. The number of clusters is determined adaptively and the parameters of the Gaussian mixture are calculated by Expectation Maximization algorithm. Finally, a layout analysis is used to finalize the detected regions as text regions.

The simulation results promise considerable improvement in both false alarms and errors seen.

In [49], three different text extraction methods are proposed. First method is stated as *Gray-level Image Analysis*, which is based on gray level stretching and binarization by average intensity of image, whereas the second method is based on Split-and-Merge and the last method is the combination of these two. Preprocessing stage of the first method consists of contrast stretching median filter, high-pass filtering and an opening operation. Lastly, edge image is extracted by a Laplacian operator. After this preprocessing step, mean intensity-based binarization is followed by connected component extraction to segment text regions. In the Split and Merge Process, a segment of an image is split into four regions according to region intensity, then in the merging step regions are merged again according to their intensities. After this step, a size pruning and a dilation process is applied. Finally, text extraction step is applied to obtain rectangular regions. The third approach utilized from the observation of Gray-level Image Analysis' sensitivity to horizontal and vertical lines in addition to Split and Merge Analysis' sensitivity to color similarity and their robustness to each others weaknesses.

In another study [55], Kim et al. has proposed a scene text extraction system, in which low level image features, such as color continuity, gray level variation and color variance, are combined hierarchically with a high-level feature, where these high-level features are extracted by examining text stroke by using a multi-resolution wavelet transform on image regions selected by the low-level features. Finally, the feature vector obtained from the last step is fed to a Support Vector Machine for verification. More thoroughly, system obtains 3 extraction results from low-level features. Based on these features, it is aimed to increase robustness against lighting variations, directional changes and perspective distortions. Low-level features are combined by the 80% rule [55] and if a candidate region is overlapped by two of the results, no verification is needed; in any other case to

verify the region, SVM with key-stroke analysis is incorporated. The verification is achieved by dividing regions as 16x16 overlapping block and classifying each block as text or non-text. Although detection results are not reported in the study, affine rectification has considerably improved character recognition rate of the overall system.

In [57], a scene text extraction system for language translation on a handheld device is proposed. The text extraction system integrates a symmetric neighborhood filter to enhance image, a hierarchical connected components to segment characters from background in the scene and a wavelet transformation to verify text regions. Due to the noise, illumination changes and motion of camera a degree of degradation occurs on the acquired images and in this approach image is first enhanced by a symmetric neighborhood filter (SNF) filter [98] in order to overcome these degradations. Then, a hierarchical connected component analysis follows. After obtaining connected components, a binarization is applied. Moreover, by the introduction of a text-similarity measure based on Discrete Wavelet coefficients, the regions are verified. After the verification step, text-line boundaries are computed by simple horizontal projection profile. Finally, off-the-shelf OCR software is used to recognize characters. Among 47 words contained in 36 images 41 words are successfully recognized.

In a different study, in order to capture the parts-based nature of the text, Zhang et al in [45] have proposed a text extraction scheme, in which scene text is detected by using a higher-order MRF. In their work, a region adjacency graph (RAG) is used to model relations of parts and properties of the parts. The regions are detected by a mean-shift segmentation algorithm [70], which incorporates color space of the image. An edge of the graph is only established, if it is close enough and the values of minimum distance threshold allow three consecutive characters form a three-clique. The features for detection of scene text are one-node features and three-node features. One-node feature only consists of aspect ratio modeled as Gaussian probability distribution functions, whereas three node features consists

of minimum angle, consistency of region inter-distance, maximum color distance and height consistency of the character. Maximum Likelihood estimation is used for the inference of higher order MRF. Furthermore, a variational method, in the form of belief propagation [71], is developed for inference in the higher order model. In the overall miss rate is calculated as 33% whereas hit rate is obtained as 67% amongst the 20 test images. However, this method is quite promising due to the method's capacity to capture the multi-part nature of scene-text by a statistical approach.

## *3.3 Related Work Focusing on Challenges of Scene-text*

### *3.3.1 Perspective Projection Distortion*

Since most of the OCR systems are not perspective intolerant, in other words do not handle perspective distortions, a perspective rectification is required for the accurate recognition of any scene-text. Assuming that text is on a plane and at an angle relative to imaging plane, the distortion can be described as a homography between the imaging plane and the text plane. This transformation – from text plane to imaging plane – can be modeled by a 3×3 matrix, in which calculation of 8 parameters – excluding the normalization factor – are required to successfully inverse the effects of perspective.

In [77], Hsieh et al. introduce a morphology-based method for detecting license plates from cluttered scenes. In their study, images containing license plates are captured in various camera orientations producing distortion on license plates. Next, a simple linear approximation is used to rectify the inclined plate to increase the recognition performance. Since the perspective distortion in their case is weak, such a crude method shows considerable performance.

In another study, Jung et al. try to estimate text region's coordinate system by incorporating the bounding lines of text region and eight parameter

correspondences. In their work [78], the text regions are assumed to be rectangular and contained on planar surfaces with homogeneous background. The warping parameters are estimated by establishing correspondence between the imaging and text planes via four vertices, which are obtained from the intersection of bounding lines. However, these assumptions coupled with their estimation method for correspondence parameters, make their approach unattractive. Perspective is stronger than their assumption in most of the cases. And line-fitting using straight forward least square method can be another disadvantage.

Assumption of text regions residing on planar surfaces is quite common in the scene-text detection literature. Gandhi, Kasturi and Antani propose a scene-text detection method based on the planar motion segmentation [79]. Motion model parameters of planar text surfaces are estimated by using gradient-based methods from a sequence of images. Moreover, plane equations are calculated from motion models of multiple surfaces. Furthermore, perspective distortions of detected planar surfaces are corrected by using normal vectors of these planes, which are calculated from plane equations. However, this method [79] assumes that camera calibration parameters are known which is not the case most of the time; and also their method requires multiple images and therefore would not work on single images.

Chen et al. report in [55] a sign text detection method which incorporates an affine restoration in order to improve both detection and recognition performances. In their study, following the coarse detection of text, parallel lines in the 3D space are used to compute the normal vector of sign plane. These parallel lines are extracted from sign's rectangle frame, direction of text (which is estimated by Hough transform) and from the first and the last character in the text row. After employing affine restoration by incorporating this normal vector furthermore, a B-Spline based interpolation is employed. It should be noted that this method extensively relies on the assumption of bounding frames around signs.

Under the assumption of mild perspective, Clark and Mirmehdi [81] establish a rectification method that does not use vanishing points or any perspective information. Instead, they estimate and recover perspective effects on text using the quadrilateral bounding box, which represents a rectangular box on the text plane. They also make use of bilinear interpolation to decide which source pixel corresponds to each destination pixel. In a later work [82] from these two authors, they introduce vanishing points in order to handle distortions under strong perspective. In order to estimate the horizontal vanishing point, they use an extended 2D projection profile in which separation of lines are maximized in a paragraph. In their study, they find the vertical vanishing point from the alignment of paragraph or the type of justification. While achieving this goal, they assume that text regions contain a number of text lines. However, in the real world scenes, text regions have most of the time single lines.

Myers et al. in [80] illustrate that some of the eight parameters in the perspective transformation are not required to be calculated. Myers point out that the parameters introducing translations in $x$- and $y$-directions and scaling in both dimensions are already handled by typical OCR systems. Furthermore, as they indicate, skew can be computed from traditional page analysis. Therefore, they conclude that there are three critical parameters that produce distortions which are problematic for OCR systems; namely, shearing and two perspective foreshortening parameters. In their approach firstly, basic features, such as base line, top line of text and the dominant vertical direction of characters are extracted. Using base and top lines horizontal vanishing point is found. Next, assuming weak perspective at $y$-direction, the vertical vanishing point is located at infinity. Finally rectification is achieved by applying the reduced transformation obtained from the two vanishing points. Reported results show significant improvements in OCR performance hence validating the assumptions made. However, capture resolutions below SVGA (800×600) can still induce problems

on character recognition even if perfect rectification is applied since characters tend to merge and consequently hard to segment in low-resolutions.

### *3.3.2 Non-planar Surfaces:*

Most of the related studies on non-planar surfaces issue in the literature are focused on the analysis of documents with curved and/or cylindrical surfaces. A number of works, such as [83], [87], take the advantage of deviations of the distorted text lines from straight lines; hence, they employ a curve fitting approach to correct warping effect. In another approach [88],[89],[91], the assumption of a cylindrical model is proposed, since the deformation is assumed to be formed by spine of a book; hence, this assumption is well grounded. However, in general camera-based scene-text detection and recognition, assumption of cylindrical surface is invalid. In that case, text surfaces might take the form of arbitrary shapes, which on contrary to perspective distortion produces a nonlinear transformation to compensate. Although the methods in [84],[85],[86] deal with arbitrary surfaces by employing a mesh model, in these methods specialized equipments – such as light grid projector – are used to extract 3D surface information. For a successful scene-text detection and recognition system, it is beneficial that system is mobile, low-cost and simple as pointed out in [66]. Thus, using additional equipment would increase cost and complexity, while decreasing mobility. Ulges et al. in [90] describes a system that determines the arbitrary shape of a page from multiple images captured by using an off-the-shelf digital camera. In their work, general purpose vision methods, such as epipolar geometry, feature matching and disparity, are employed. Moreover, surfaces of pages are modeled with a triangular mesh structure by using disparity information extracted from multiple images. Finally, an image dewarping is applied to correct the nonlinear transformation. Although this work is a right step towards constructing a mobile and low-cost scene-text detection and recognition system, ideally it is desirable to estimate the shape of the text surfaces directly from a single image.

### *3.3.3 Capture Issues*

Since camera captured image resolutions is often below 150 dpi, the characters in a scene-text can be quite small to be recognized by typical OCR systems. An enhancement of resolution can be beneficial to the overall performance of the system. It is also worth mentioning that for this kind of enhancement processing of multiframe sequences is necessary and advantageous compared to single frame processing methods, such as bilinear interpolation. When multiple frames are available for processing, super-resolution is often the first tool to be considered for enhancing such images.

In [92], Li and Doermann utilize projection onto convex sets (POCs) approach to obtain higher resolution and deblured scene text image. Capel and Zisserman in [93] reported an evaluation of four estimators for super-resolution enhancement of text; namely a maximum likelihood (ML) estimator, a maximum a posteriori (MAP) estimator, an iterative back-projection method and an estimator regularized by using the Total Variation norm. The evaluation demonstrated the superiority of Total Variation norm based super-resolution and MAP estimator.

Another approach proposed by Baker and Kanade in [95], introduces example-based priors incorporated with a Bayesian framework. In this work, authors first show that any smoothness constraints causes over smoothing for large magnification factors, even though there are more than enough low-resolution images. Furthermore, they develop a method, which learns local features from training set and enhances these features in low resolution images to obtain better super-resolution results. With this approach, they achieve better results on non-text imagery (i.e. images containing faces) and on document images. However, they lack of presenting any performance evaluation of character recognition with the proposed method.

Donaldson and Myers propose a Bayesian scene-text super-resolution algorithm, which uses a text-specific bimodal prior in [94]. Bimodal prior is inferred from observing that for most of the text in scenes character pixels tend to cluster around one value, while background pixels cluster around another intensity value. In their study, the bimodality is introduced by using an exponential distribution with a fourth-order polynomial exponent. It is reported that, bimodal prior improved the recognition rate compared to piecewise smoothness prior and also the recognition of short characters compared to Bicubic interpolation. Although bimodal prior results are promising, the illumination variations on the text region can severe this assumption. A simple extension to this method can be multimodal prior which would model the noise and illumination effects on text regions.

Even though the following study aims of enhancing only document images, it is still worth mentioning in this context. Traditionally multiple observations, each representing different sampling of the same scene, are needed for super-resolution processing; however in [96], Dalley et al. propose a single-frame Bayesian-based super-resolution technique which utilizes training samples of character patches. In a nutshell, this technique first compiles a training database from low- and high-resolution image patches of the same character. In the training stage the probability of having the exact grayscale value in high-resolution image from the given low-resolution patch is estimated. After this stage, the super-resolution is applied by deciding the output pixel values from the local estimate of aforementioned probability corresponding to the low-resolution patches. Although quite promising results are shown, analysis of character recognition performance is missing from the aforementioned study.

Apart from super-resolution approaches, another enhancement procedure that can improve the performance of both detection and recognition would be boosting the text-like features in the image, in addition to decreasing blurring on text regions as much as possible. Chen et al. in [97] observe that text characters contain substructures, which exhibit stripe-like properties, and in addition these stripes

have little variation within a word. In their work [97], they derive a new form of filter from Gabor kernels to estimate scales of these features. Afterwards, the contrast of text is increased by enhancing only the edges of stripes with specific scale representing the text. This results a smoothed background while text edges are preserved and improved as well. Furthermore, they report better performance results in the binarization step by using this approach. In another study [10], bilinear interpolation extended by introducing the temporal information present in the video. In which, localized temporal mean and standard deviation is incorporated. Haritaoglu and Haritaoglu [98] propose a fast image enhancement method by using a symmetric neighborhood filter which can sharpen the blurred transitions while preserving boundary edges. All these methods yield some improvement on text images, although the resulting improvement is not capable of completely compensating issues introduced by low-resolution and noisy capture such as touching characters.

## *3.4 Scene-text Localization Results*

Some experiments are conducted on ICDAR 2003 Text Locating Competition Test Set, which contains 544 various indoor/outdoor scene images. Another important point to be noted is that most of the images in the set are captured as high resolution, larger than 800×600 and in order to decrease the computation time, all these images are down-sampled to 352×288 resolution before any processing takes place.

For the estimation of necessary parameters, such as covariance matrices, mean vectors and etc., 446 frames containing overlay-text regions captured from 21 different Turkish television channels are used. The uniform block size is selected as 8×8 pixels and 27 dimensional feature vectors that are constructed from 64 DCT coefficients are used for the representation of each image block.

In Figure 26 and Figure 27, sample localization results are presented. In these images, blue rectangles depict ground-truth regions, whereas green boundaries represent the localization results due to the proposed scheme.


**Figure 26: Localization results (Green boxes indicate results due to the proposed algorithm whereas blue rectangles are the ground-truth).**

Overall text locating results on the competition data set is calculated as 71.44% for the Recall and 44.32% for the Precision, which are tabulated in Table 8. Moreover, in Table 8, ICDAR 2003 Text-Locating Competition results [99] are also presented for comparison. Evaluation method used in the construction of the following table is described elaborately in [99].

**Table 8: Performance comparison of different systems in [99].**

| ***System*** | ***Precision*** | ***Recall*** | ***f*** |
|---|---|---|---|
| *Proposed* | *0.44* | *0.71* | *0.54* |
| Ashida | 0.55 | 0.46 | 0.50 |
| HWDavid | 0.44 | 0.46 | 0.45 |
| Wolf | 0.30 | 0.44 | 0.35 |
| Todoran | 0.19 | 0.18 | 0.18 |
| Full | 0.10 | 0.08 | 0.08 |

Although, the proposed method outperforms other entries in terms of hit-rate performance, precision performance does not yield the same success. One of the many reasons of this low precision can be explained by the training set that is utilized. As it has already noted before, during the training process, images that contain overlay-text (not scene-text) regions are utilized which, causes substantial increase in the false-alarms. Furthermore, noting that dominant character-resembling structures in cityscape images produces most of the false-alarms observed in results, the abundant number of cityscape images in competition data set effects those precision values in a drastic manner.

Another point worth to mention is about the utilized word segmentation method; it does not handle perspective distortion that most of the scene-text contains; instead it is assumed that text regions are perfectly aligned with the imaging plane. Moreover, down-sampling of the images due to dealing with computational complexity causes an information loss.


**Figure 27: Scene-text localization results.**

## 3.5 Perspective Rectification for Recognition of Scene-text on Signs

As described elaborately in Section 3.1.1, perspective projection distortion has a significant performance diminishing effect. Although complete compensation of this effect depends on the circumstances, performance of scene-text recognition can be improved to a greater extend by using a rectification process. In this effort, mainly rectification of text on signs is addressed. Moreover, bounding frames around signs are assumed and utilized in order to rectify the scene-text. Additionally, weak perspective projection is assumed as in [80] and [81], since generally the variation in the depth of scene-text plane is small compared with the distance between camera and the plane. In the following subsections necessary preliminaries and the proposed method is described. Furthermore, algorithm is tested against real world data.

### 3.5.1 Preliminaries

The transformation from a three-dimensional scene to a two-dimensional image is called a *projection*. Furthermore, if the points used in the image formation are coplanar – e.g. 3D points on a sign – this transformation can be described by a *homography (or projective transformation).* )This homography can be undone by applying the inverse transformation to the each pixel in the observed image as described in [102] and [80];

$$p' = Hp \qquad (2.25)$$

In Equation (2.25), the homogeneous coordinates $p = \begin{bmatrix} x & y & 1 \end{bmatrix}^t$ in the observed image is mapped to homogenous coordinates $p' = \begin{bmatrix} kx' & ky' & k \end{bmatrix}^t$ in the rectified image by the 3×3 linear transformation *H*. This transformation can further be

decomposed into a chain of transformations [102]; namely, a 2D Euclidean transformation $H_E$, an affine transformation $H_A$ and the projective transformation $H_P$ (2.25).

$$p' = H_E H_A H_P p \quad , \tag{2.25}$$

$$H_E = \begin{bmatrix} s\cos\theta & s\sin\theta & t_x \\ -s\sin\theta & s\cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \quad H_A = \begin{bmatrix} 1/b & -a/b & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad H_P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ l_x & l_y & 1 \end{bmatrix} .$$

Finally point-to-point correspondence can be used to compute the aforementioned homography. However, before describing the computation of projective transformation, significance of parameters in (2.25) should be mentioned. $t_x$ and $t_y$ in $H_E$ are the translation parameters on both $x$ and $y$-direction, whereas $\theta$ and $s$ are the rotation and scale parameters respectively. Likewise, affine transformation $H_A$ has a scaling parameter of one axis relative to the other namely, $b$ and a shear parameter, $a$. Lastly, $l_x$ and $l_y$ in $H_P$ represent the perspective foreshortenings along the directions $x$ and $y$.

### *3.5.2 Proposed Approach*

As mentioned before some of the distortions introduced by the perspective projection are handled in the character recognition level, such as the translation and scaling in $x$ and $y$ directions. Noting this previous remark, the number of point-to-point correspondences necessary for the computation of projective transformation can be decreased from 8 to 4 correspondences by excluding aforementioned parameters. Two of these correspondences can be extracted from the horizontal and vertical vanishing points in the observed image, which are mapped to the points at infinity in the rectified image. At this point, it should be noted that although horizontal vanishing point can be described by $\begin{bmatrix} x & y & 1 \end{bmatrix}^t$, because of the weak perspective projection assumption vertical lines in the

observed plane are only affected from shear and also, parallel lines on the 3D scene are still parallel. As a result vertical vanishing point is located at $[\alpha \quad 1 \quad 0]^t$, with the shear factor $\alpha$. The other correspondence points can be selected arbitrarily such as the corners of the frame around the plane that is, $[0 \quad 0 \quad 1] \rightarrow [0 \quad 0 \quad 1]$ and $[1 \quad 1 \quad 1] \rightarrow [1 \quad 1 \quad 1]$. With these correspondences, the inverse transformation can be computed as in [80];

$$H = \begin{bmatrix} -(x-y)(x-\alpha y+\alpha-1) & \alpha(x-y)(x-\alpha y+\alpha-1) & 0 \\ -y(\alpha-1)(x-\alpha y+\alpha-1) & x(\alpha-1)(x-\alpha y+\alpha-1) & 0 \\ -(\alpha-1)(x-y) & \alpha(\alpha-1)(x-y) & (\alpha-1)(x-y)(x-\alpha y) \end{bmatrix}$$

*(2.25)*

Computation of vanishing points are carried out by first extracting edges of the sign by using a robust edge detection method e.g. Sobel operator. Afterwards, Generalized Hough Transform is utilized in order to extract four sides of the frame around the sign. While extracting sides from Hough Transform, as a means to discard false candidates of lines, angle and magnitude thresholds are introduced. Finally, horizontal vanishing point is found from the intersection points of top line and baseline whereas; other two sides are used to locate the vertical vanishing point.

### *3.5.3 Preliminary Simulations*

As a preliminary study of improvement achieved by the perspective rectification, recognition performance with and without rectification is compared. For this purpose a single word plastered on a wall is captured in various angles with the size of 352×288 and recognition results on each image is recorded for both cases.

It should be noted that for the sake of simplicity positions of corners essential for the inverse transformation are manually identified in each test image. In Figure 29 set of the test images are shown whereas Figure 28 depicts performance results. As it can be seen from the graph, rectification process improves the recognition performance significantly in angles between 30° and 60°.

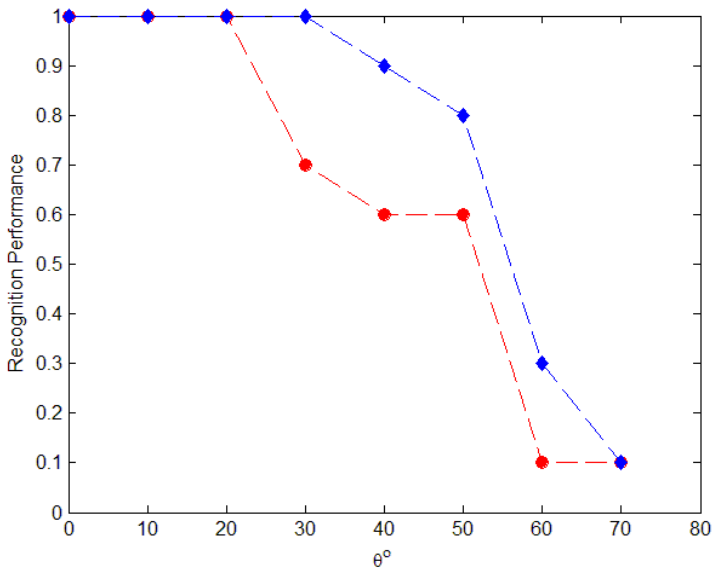



**Figure 28: Comparison of Recognition Performance with/without Rectification.**


**Figure 29: Test images captured in different angles. For each image angle varies in increments of 10°.**

### *3.5.4 Simulation Results*

In Figure 31 and Figure 32, typical examples for rectification process with its intermediate steps are depicted. Moreover, word recognition results can be observed from Figure 30 and additionally more elaborate results are shown in Table 9 for Figure 31. As it can be seen from these evidences, rectification method proposed can effectively rectify the perspective projection distortion and also increase the recognition performance of the overall system. However, since the overall rectification method relies on the extraction of sign's frame any errors on that causes whole rectification process to fail (Figure 33).

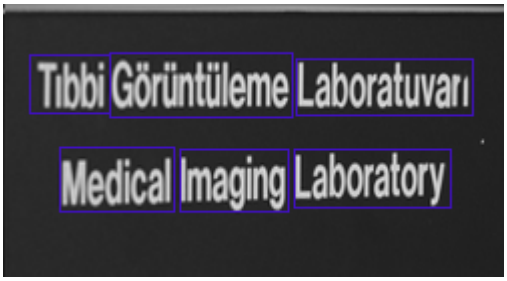


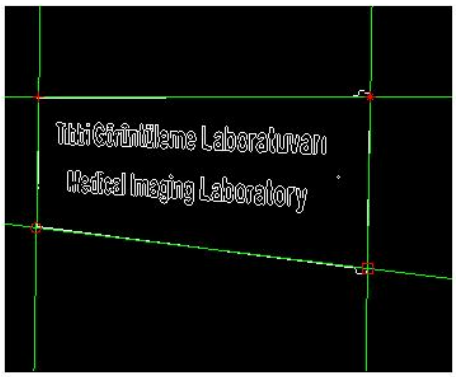

**Figure 30: Word recognition results obtained from image rectified image; "TIBBI", "GDRONTUIEME", "LABORATUVARI", "MEDICAI", "IMAOJNG", "LABORATORY".**

**Table 9: Word recognition results obtained from Figure 31-a. NR: Without Rectification, WR: With Rectification Process, WRP: With Rectification and Post-Processing**

| ***Word*** | ***NR*** | ***WR*** | ***WRP*** |
|---|---|---|---|
| TIBBI | TCR | TIBBI | TIBBI |
| GORUNTULEME | GGGIHNE | GDRONTUIEME | GORUNTULEME |
| LABORATUVARI | UUVTOARI | LABORATUVARI | LABORATUVARI |
| MEDICAL | MHNN | MEDICAI | MEDICAL |
| IMAGING | INAGI | IMAOJNG | IMANLI |
| LABORATORY | LABORATON | LABORATORY | LABORATORY |


(a)

(b)

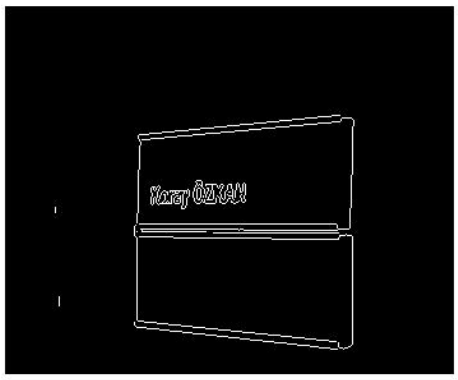


(c)

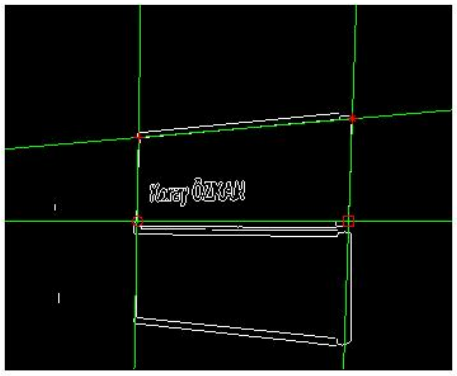

(d)

**Figure 31: (a) Processed image, (b) Edge Map, (c) Extracted Lines, (d) Rectification result.**


(a)



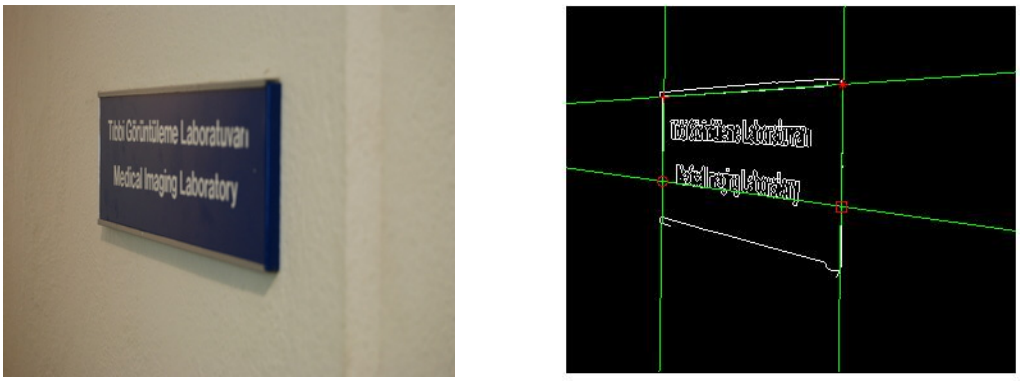
(b)



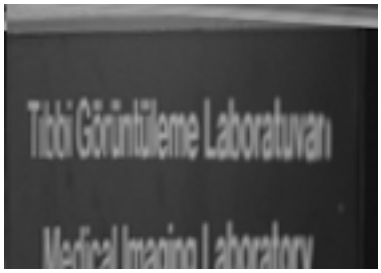
(c)


(d)

**Figure 32: (a) Processed image, (b) Edge Map, (c) Extracted Lines, (d) Rectification result. Word recognition result obtained after rectification; KORMY and OZDEN**




(a) (b)




(c)

**Figure 33: Result obtained when text-plane extraction is failed. (a) Processed image, (b) Extracted Lines, (c) Rectification result.**

# CHAPTER 4

# SUMMARY, DISCUSSIONS AND FUTURE WORKS

## *4.1 Summary of the Thesis*

In this research effort, problem of text information extraction from digital media is examined. Primary building blocks of a Video OCR system are identified, investigated and engineered during this endeavor and a detailed analysis of each building block is discussed throughout the dissertation. The body of work is divided into two main parts according to the inherent properties of the text in video.

In Chapter 2, detection, localization and recognition of overlay-text is thoroughly analyzed. After an extensive literature survey on the subject of detection is provided; framework for an overlay-text localization system is proposed. The framework consists of feature extraction, classification and a bounding-box extraction steps. For feature extraction, several methods are suggested and additionally, a novel DCT-based feature vector extraction approach is presented. Moreover, candidate methods, such as Learning Vector Quantization, for the classification of these feature vectors are introduced. Finally, a comparative analysis of both feature extraction and classification methods is presented. Furthermore, Markov Random Field based text/non-text classification approach is discussed in order to incorporate the multi-part nature of text in digital media.

Furthermore, the performance of this approach is compared with the naïve Bayesian approach. In the proposed framework, classification results are further processed to extract text regions as word-by-word. For this purpose, a novel bounding-box extraction method is also introduced in this chapter. Additionally, the experimental results of overall localization system are provided on a challenging and vast data set.

After text detection is analyzed, in the accompanying sections of the chapter, firstly, components of a complete recognition system is introduced and discussed in detail. For the enhancement block of this recognition system, two well-known interpolation techniques, namely Bilinear and Bicubic, are compared. Afterwards, segmentation of text/background is visited by investigating some commonly used thresholding methods. Furthermore, a multi-hypothesis based segmentation method is proposed, instead of simple binarization techniques and recognition performance of this approach is compared with the conventional thresholding methods. Towards a complete recognition module, a character recognition method is adopted from a well-established *handprint* OCR system. After introducing the method for extracting single characters from image data, Karhunen Loève transform is presented for the feature vector extraction. Moreover, for the classification of each feature vector of characters, Multi-Layer Perceptron is presented, by not only discussing the topology and properties of the network but also the training process and test results. Lastly, a novel post-recognition spelling improvement method is designed based on ranking of dictionary-generated candidates for recognition results. For the ranking mechanism in this structure, confidence values obtained from the character classification is utilized.

Chapter 3 discusses the scene-text localization and recognition problem. At first major challenges complicating the construction of a complete scene-text recognition system are identified. A detailed literature survey focusing specifically on the determined challenges is also provided in this chapter. Furthermore, for the scene-text detection problem, a thorough analysis of related

work is presented with discussion points. Afterwards, localization performance of the previously discussed framework for Video OCR system is obtained on a standardized ICDAR Text-Locating Competition dataset. Furthermore, the results are compared with other contestant systems. Lastly, a novel method for compensation of perspective projection distortion for scene text data is proposed. This method specifically applies on the domain of scene-text that is present on signs and it utilizes Generalized Hough Transform. Moreover, benefits of a rectification method are demonstrated by presenting the recognition results.

## *4.2 Discussions and Future Work*

Throughout the investigation of literature and analysis of the domain by simulations, it has been observed that accurate representation of text needs a lot of attention during the system design. Thus, different methods have been compared in order to find the appropriate feature extraction method for the solution of the problem. However, it is also inferred from experiments that a fusion of different representation methods can enhance the performance of classification, which remains as a future work.

It should be noted that in this study it is also intended to design a system not only accurate and robust but also simple from the computational complexity perspective. Hence, the system is tried to be build on computationally efficient methods, such as the naïve Bayesian classifier. Furthermore, multi-part nature of text is exploited in order to improve text/non-text classification performance by incorporating Markov Random Fields. Although favorable performance gains are reached, considering more sophisticated energy models is an important direction to investigate. Another equally important, future investigation point would be the minimization methods used for the solution of energy model, other than Iterated Conditional Mode, since during the simulations it has been observed that ICM's deterministic nature and dependence to initial conditions effects performance.

By introducing bounding-box extraction block to the system, text detection results are further processed to obtain proper text regions for recognition. This way, the amount of unnecessary data to be processed by recognition module is lowered. The simulations on a test set which spans a large unconstrained generic television broadcast indicate considerable maturity of the proposed Video OCR system. Especially the results on stationary, and high-contrast text regions that are not affected by noise, motion blur and capture issues, are observed to yield much better results. However, utilization of this system in real applications, such as video indexing, forces to carefully design supplementary modules to overcome aforementioned and many more difficulties, which is left as another future work.

From the perspective of recognition, complex background, noise and several grayscale distribution modes present on the localized text regions make the optimal binarization of text regions problematic. However, in this work, character recognition performance is improved by introducing multi-hypothesis approach based on the K-Means clustering algorithm. Furthermore, problems caused by the low resolution nature of the video data is tried to be reduced by adopting a character segmentation method that is used in a handprint recognition system. With these design modifications and tunings, it is once again proven that for successful Video OCR application, typical off-the-shelf OCR packages are not adequate.

At this point, it should be also noted that any sort of binarization process on the image data results a considerable reduction in the amount of valuable information necessary for recognition. Although, multi-hypothesis approach strives to improve this drawback, and succeeds to some degree, gray-scale based character recognition is an important direction to be considered for the solution of aforementioned difficulty. Not only the single character recognition performance but also character segmentation can be improved by renovating character recognition module so that no binarization is required.

Scene-text localization, given the nature of the domain, still remains as a challenging topic. Although, in the experimental setup the proposed method shows a considerable performance, for real-life applications, further development is necessary. During the simulations, it is also observed that the complex scene issue and the non-planar text surfaces are major bottlenecks for localization of scene-text. In this respect, the method proposed for the rectification of perspective distortion relies heavily on the assumption of text on signs. Moreover, it is expected that the frame of the signs are easily extracted, which may not be the case for more generic data. A complete rectification system would also incorporate the direction of text. In addition, arbitrary shape extraction methods can also be incorporated to compensate the perspective distortion as well as to solve non-planar text surfaces problem.

# REFERENCES


[1] Huiping Li; Doermann, D.; Kia, O., "Automatic Text Detection and Tracking in Digital Video," *IEEE Transactions on Image Processing,* vol.9, no.1, pp.147-156, Jan 2000.

[2] Huiping Li; Doermann, D., "A Video Text Detection System Based On Automated Training," *Proceeding of 15th International Conference on Pattern Recognition,* vol.2, no., pp.223-226 vol.2, 2000.

[3] Lienhart, R., "Automatic Text Recognition for Video Indexing," In *Proceedings of the Fourth ACM international Conference on Multimedia*, 1996.

[4] Yu Zhong; Karu, K.; Jain, A.K., "Locating Text In Complex Color Images," *Proceedings of the Third International Conference on Document Analysis and Recognition,* vol.1, no., pp.146-149 vol.1, 14-16 Aug 1995.

[5] Chaddha N., Sharma R., Agrawal A., and Gupta A., "Text Segmentation in Mixed–Mode Images," *In 28th Asilomar Conference on Signals, Systems and Computers,* pp. 1356-1361, October 1995.

[6] Dimitrova N., Agnihotri L., Dorai C., Bolle R., "MPEG-7 Videotext Description Scheme For Superimposed Text In Images And Video," *Signal Processing: Image Communication,* vol. 16, no. 1, pp. 137-155, 2000.

[7] Smith, M.A.; Kanade, T., "Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol., no., pp.775-781, 17-19 Jun 1997.

[8] Crandall, D.; Antani, S.; Kasturi, R., "Extraction of Special Effects Caption Text Events from Digital Video," *International Journal on Document Analysis and Recognition,* no.pp. 138-157 vol. 5, 2003.

[9] Mariano, V.Y.; Kasturi, R., "Locating Uniform-Colored Text in Video Frames," *Proceedings of 15th International Conference on Pattern Recognition,* vol.4, no., pp.539-542 vol.4, 2000.

[10] Wolf, C.; Jolion, J.-M.; Chassaing, F., "Text Localization, Enhancement and Binarization in Multimedia Documents," *Proceedings of 16th International Conference on Pattern Recognition*, vol.2, no., pp. 1037-1040 vol.2, 2002.

[11] LeBourgeois, F., "Robust Multifont OCR System from Gray Level Images," *Proceedings of the Fourth International Conference on Document Analysis and Recognition,* vol.1, no., pp.1-5 vol.1, 18-20 Aug 1997.

[12] Shin, C.S.; Kim, K.I.; Park, M.H.; Kim, H.J., "Support Vector Machine-Based Text Detection in Digital Video," *Proceedings of the IEEE Signal Processing Society Workshop Neural Networks for Signal Processing X,* vol.2, no., pp.634-641 vol.2, 2000.

[13] Kwang In Kim; Keechul Jung; Jin Hyung Kim, "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,* vol.25, no.12, pp. 1631-1639, Dec. 2003.

[14] Lienhart, R. and Stuber, F., "Automatic Text Recognition in Digital Videos," *In Image and Video Processing IV 1996, Proc.* SPIE 2666-20, Jan. 1996.

[15] Gllavata, J.; Ewerth, R.; Freisleben, B., "Text Detection in Images Based on Unsupervised Classification of High-Frequency Wavelet Coefficients," *Proceedings of the 17th International Conference on Pattern Recognition,* vol.1, no., pp. 425-428 Vol.1, 23-26 Aug. 2004.

[16] Lyu, M.R.; Jiqiang Song; Min Cai, "A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction," *IEEE Transactions on Circuits and Systems for Video Technology,* vol.15, no.2, pp. 243-255, Feb. 2005.

[17] Min Cai; Jiqiang Song; Lyu, M.R., "A New Approach for Video Text Detection," *Proceedings of International Conference on Image Processing.* vol.1, no., pp. I-117-I-120 vol.1, 2002.

[18] Yu Zhong; Hongjiang Zhang; Jain, A.K., "Automatic Caption Localization in Compressed Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.22, no.4, pp.385-392, Apr 2000.

[19] Tseng, B.L.; Ching-Yung Lin; Dongqing Zhang; Smith, J.R., "Improved Text Overlay Detection in Videos Using a Fusion-Based Classifier," *Proceedings of International Conference on Multimedia and Expo,* vol.3, no., pp. III-473-6 vol.3, 6-9 July 2003.

[20] Zhang, D.Q.; Tseng, B.L.; Chang, S.-F., "Accurate Overlay Text Extraction for Digital Video Analysis," *Proceedings of International Conference on Information Technology: Research and Education,* vol., no., pp. 233-237, 11-13 Aug. 2003.

[21] Jae-Chang Shim; Dorai, C.; Bolle, R., "Automatic Text Extraction from Video for Content-Based Annotation and Retrieval," *Proceedings of Fourteenth International Conference on Pattern Recognition,* vol.1, no., pp.618-620 vol.1, 16-20 Aug 1998.

[22] Choi, Y.; Chee Sun Won; Yong Man Ro; Manjunath, B. S.; "Texture Descriptors", *Introduction to MPEG-7: Multimedia Content Description Interface, Wiley,* 2002.

[23] Park, D. K., Jeon, Y. S., and Won, C. S., "Efficient Use Of Local Edge Histogram Descriptor," *In Proceedings of the 2000 ACM Workshops on Multimedia* pp. 51-54, October 30 - November 03, 2000.

[24] S. J. Park, D. K. Park, C. S. Won, "Core Experiments on MPEG-7 Edge Histogram Descriptor," MPEG document M5984, Geneva, May, 2000.

[25] Yong Man Ro, Munchurl Kim, Ho Kyung Kang, B.S. Manjunath, and Jinwoong Kim, "MPEG-7 Homogeneous Texture Descriptor" *ETRI Journal,* vol.23, no.2, June 2001, pp.41-51.

[26] Kohonen, T.; Kangas, J.; Laaksonen, J.; Torkkola, K., “LVQPAK: A Software Package for the Correct Application of Learning Vector Quantization Algorithms,” *International Joint Conference on Neural Networks,* vol.1, no., pp.725-730 vol.1, 7-11 Jun 1992.

[27] Huawu Deng; Clausi, D.A., "Unsupervised Image Segmentation Using a Simple MRF Model with a New Implementation Scheme," *Proceedings of the 17th International Conference on Pattern Recognition,* vol.2, no., pp. 691-694 Vol.2, 23-26 Aug. 2004.

[28] Melas, D.E.; Wilson, S.P., “Double Markov Random Fields and Bayesian Image Segmentation,” *IEEE Transactions on Signal Processing,* vol.50, no.2, pp.357-365, Feb 2002.

[29] Ioffe, S.; Forsyth, D.A., “Probabilistic Methods for Finding People,” *International Journal of Computer Vision,* vol. 43, pp.45-68, June 2001.

[30] Felzenszwalb, P.F.; Huttenlocher, D.P., “Efficient Matching Of Pictorial Structures,” *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* vol.2, no., pp.66-73 vol.2, 2000.

[31] S. Z. Li, “Markov Random Field Modeling in Computer Vision” *Springer-Verlag,* 2001.

[32] Jain, R.; Kasturi, R.; Schunk, B.G.; “Machine Vision,” *McGraw-Hill, International Editions,* 1995.

[33] Garris, M. D.; Blue, J. L.; Candela, G. T.; Grother, P. J. ; Janet, S. A.; Wilson, C. L., “NIST Form-Based Handprint Recognition System (Release 2.0),” *Technical Report NISTIR 5959*, January 1997.

[34] Keys, R., “Cubic Convolution Interpolation for Digital Image Processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.29, no.6, pp. 1153-1160, Dec 1981.

[35] Sankur, B.; Sezgin, M., “A Survey Over Image Thresholding Techniques and Quantitative Performance Evaluation”, *Journal of Electronic Imaging*, 13(1), 146-165, January, 2004.

[36] Datong Chen; Odobez, J.-M.; Bourlard, H., “Text Segmentation and Recognition in Complex Background Based on Markov Random Field,” *Proceedings of 16th International Conference on Pattern Recognition,* vol.4, no., pp. 227-230 vol.4, 2002.

[37] Duda, R. O.; Hart, P. E.; Stork, D. G.; “Pattern Classification”, Second Edition, *Wiley*, 2001.

[38] Rocha, J.; Sakoda, B.; Zhou, J.; Pavlidis, T.; “Deferred Interpretation of Grayscale Saddle Features for Recognition of Touching and Broken Characters,” *Proceedings of Document Recognition*, SPIE, vol. 2181, San Jose, CA, pp. 342-350, February 1994.

[39] Matan, O.; Burges, C. J. C.; LeCun Y.; Denker, J. S.; “Multi-Digit Recognition Using a Space Displacement Neural Network,” *Advances in Neural Information Processing Systems 4*, Morgan Kaufmann, pp. 488-495, 1992.

[40] Keeler, J. and Rumelhart, D. E.; “A Self-Organizing Integrated Segmentation and Recognition Neural Net,” *Advances in Neural Information Processing Systems 4*, Morgan Kaufmann, pp. 496-503, 1992.

[41] Casey, R.G.; Lecolinet, E., "A Survey Of Methods And Strategies In Character Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.18, no.7, pp.690-706, Jul 1996.

[42] Kukich, K., "Technique for Automatically Correcting Words in Text," *ACM Comput. Surv.* Vol. 24, no. 4, 377-439, Dec. 1992.

[43] Navarro, G.; "A Guided Tour to Approximate String Matching," *ACM Comput. Surv.* Vol. 33, no. 1, pp.31-88, Mar 2001.

[44] Huiying Shen, James Coughlan, "Finding Text in Natural Scenes by Figure-Ground Segmentation," ICPR pp. 113-118, 2006.

[45] Dong-Qing Zhang, Shih-Fu Chang, "Learning to Detect Scene Text Using a Higher-Order MRF with Belief Propagation," CVPRW, pp. 101, 2004.

[46] Xiangrong Chen, Alan L. Yuille, "Detecting and Reading Text in Natural Scenes," CVPR, pp. 366-373, 2004.

[47] Renjie J.; Feihu Q.; Li X.; Guorong W., "Detecting and Segmenting Text from Natural Scenes with 2-Stage Classification," *Sixth International Conference on Intelligent Systems Design and Applications,* vol.2, no.pp.819-824, Oct. 2006.

[48] Saoi, T.; Goto, H.; Kobayashi, H., "Text Detection in Color Scene Images Based on Unsupervised Clustering of Multi-Channel Wavelet Features," *Proceedings of Eighth International Conference on Document Analysis and Recognition,* vol., no.pp. 690- 694 Vol. 2, 29 Aug.-1 Sept. 2005.

[49] JiSoo K.; SangCheol P.; SooHyung K., "Text Locating from Natural Scene Images Using Image Intensities," *Proceedings of Eighth*

*International Conference on Document Analysis and Recognition,* vol., no.pp. 655- 659 Vol. 2, 29 Aug.-1 Sept. 2005.

[50] Ezaki, N.; Bulacu, M.; Schomaker, L., "Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons," *Proceedings of the 17th International Conference on Pattern Recognition,* vol.2, no.pp. 683- 686 Vol.2, 23-26 Aug. 2004.

[51] Ezaki, N.; Kiyota, K.; Minh, B.T.; Bulacu, M.; Schomaker, L., "Improved Text-Detection Methods for a Camera-Based Text Reading System for Blind Persons," *Proceedings of Eighth International Conference on Document Analysis and Recognition*, vol., no.pp. 257- 261 Vol. 1, 29 Aug.-1 Sept. 2005.

[52] Jiang Gao; Jie Yang, "An Adaptive Algorithm for Text Detection from Natural Scenes," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol.2, no.pp. II-84- II-89 vol.2, 2001.

[53] Jie Yang; Xilin Chen; Jing Zhang; Ying Zhang; Waibel, A., "Automatic Detection and Translation of Text From Natural Scenes," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing.* vol.2, no.pp.2101-2104, 2002.

[54] Jing Zhang; Xilin Chen; Jie Yang; Waibel, A., "A PDA-based sign translator," *Proceedings of Fourth IEEE International Conference on Multimodal Interfaces,* vol., no.pp. 217- 222, 2002.

[55] Xilin Chen; Jie Yang; Jing Zhang; Waibel, A, "Automatic Detection of Signs with Affine Transformation" *Proceedings of Sixth IEEE Workshop on Applications of Computer Vision*, Vol., Iss., 2002 Pages: 32- 36.

[56] Kim, K.C.; Byun, H.R.; Song, Y.J.; Choi, Y.W.; Chi, S.Y.; Kim, K.K.; Chung, Y.K., "Scene Text Extraction in Natural Scene Images Using Hierarchical Feature Combining and Verification," *Pattern Recognition, 2004. ICPR 2004.*

[57] Haritaoglu, I., "Scene Text Extraction and Translation for Handheld Devices," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol.2, no.pp. II-408- II-413 vol.2, 2001.

[58] Xuewen Wang; Xiaoqing Ding; Changsong Liu, "Character Extraction and Recognition in Natural Scene Images," *Proceedings of Sixth International Conference on Document Analysis and Recognition,* vol., no.pp.1084-1088, 2001.

[59] Shyang-Lih Chang; Li-Shien Chen; Yun-Chung Chung; Sei-Wan Chen, "Automatic License Plate Recognition," *IEEE Transactions on Intelligent Transportation Systems,* vol.5, no.1pp. 42- 53, March 2004.

[60] Anagnostopoulos, C.N.E.; Anagnostopoulos, I.E.; Loumos, V.; Kayafas, E., "A License Plate-Recognition Algorithm for Intelligent Transportation System Applications," *IEEE Transactions on Intelligent Transportation Systems,* vol.7, no.3pp. 377- 392, Sept. 2006.

[61] Watanabe, Y.; Sono, K.; Yokomizo, K.; Okada, Y., "Translation Camera on Mobile Phone," *Proceedings of International Conference on Multimedia and Expo,* vol.2, no.pp. II- 177-80 vol.2, 6-9 July 2003.

[62] Watanabe, Y.; Okada, Y.; Yeun-Bae Kim; Takeda, T., "Translation Camera," *Proceedings of Fourteenth International Conference on Pattern Recognition,* vol.1, no.pp.613-617 vol.1, 16-20 Aug 1998.

[63] Zandifar, A.; Duraiswami, R.; Chahine, A.; Davis, L.S., "A Video Based Interface to Textual Information for the Visually Impaired," *Proceedings of Fourth IEEE International Conference on Multimodal Interfaces,* vol., no.pp. 325- 330, 2002.

[64] Wienecke, M.; Fink, G.A.; Sagerer, G., "Towards Automatic Video-Based Whiteboard Reading," *Proceedings of Seventh International Conference on Document Analysis and Recognition,* vol., no.pp. 87- 91 vol.1, 3-6 Aug. 2003.

[65] Zandifar; A., Duraiswami; R., Davis; L. S., "A Video-Based Framework for the Analysis of Presentations/Posters," *International Journal on Document Analysis and Recognition,* vol., no.pp. 178-187 vol. 7, 2005.

[66] Liang J., Doermann D., Li H., "Camera Based Analysis of Text and Documents: A Survey" *International Journal on Document Analysis and Recognition,* no.pp. 84-104 vol. 7, 2005.

[67] Shental; N., Zomet; A., Hertz; T. and Weiss; Y., "Pairwise Clustering and Graphical Models," NIPS 2003.

[68] Ohta, Y.; Kanade, T.; Sakai; T., "Color Information for Region Segmentation," *Computer Graphics and Image Processing*, 13, 222-241, 1980.

[69] Wang, L.; Pavlidis, T., "Direct Gray-Scale Extraction of Features for Character Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.15, no.10pp.1053-1067, Oct 1993.

[70] Comaniciu, D.; Meer, P., "Robust Analysis of Feature Spaces: Color Image Segmentation," *Proceedings of 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol., no.pp.750-755, 17-19 Jun 1997.

[71] Yedidia; J.S., Freeman; W.T., Weiss; Y., "Bethe Free Energies, Kikuchi Approximations, and Belief Propagation Algorithms". 2001. *MERL Cambridge Research Technical Report* TR 2001-16.

[72] Yang; J., Gao; J., Zhang; Y., Chen; X., and Waibel; A., "An Automatic Sign Recognition and Translation System," In *Proceedings of the 2001 Workshop on Perceptive User interfaces* (Orlando, Florida, November 15 - 16, 2001). PUI '01, vol. 15. ACM Press, New York, NY, 1-8.

[73] Pilu; M., Pollard; S., "A Light-Weight Text Image Processing Method For Handheld Embedded Cameras", *British Machine Vision Conference*, Sept 2002.

[74] Dumitras, T.; Lee, M.; Quinones, P.; Smailagic, A.; Siewiorek, D.; Narasimhan, P., "Eye of the Beholder: Phone-Based Text-Recognition for

the Visually-Impaired," *IEEE International Symposium on Wearable Computers,* Vol., Iss., Oct. 2006 Pages:145-146.

[75] de la Escalera, A.; Moreno, L.E.; Salichs, M.A.; Armingol, J.M., "Road Traffic Sign Detection and Classification," *IEEE Transactions on Industrial Electronics,* Vol.44, Iss.6, Dec 1997 Pages:848-859.

[76] Miura, J.; Kanda, T.; Shirai, Y., "An Active Vision System for Real-Time Traffic Sign Recognition," *Proceedings of Intelligent Transportation Systems,* Vol., Iss., 2000 Pages:52-57.

[77] Jun-Wei Hsieh; Shih-Hao Yu; Yung-Sheng Chen, "Morphology-Based License Plate Detection from Complex Scenes" *Proceedings of 16th International Conference on Pattern Recognition,* Vol.3, Iss., 2002 Pages: 176- 179 vol.3.

[78] Keechul Jung; Kwang In Kim; JungHyun Han, "Text Extraction in Real Scene Images on Planar Planes," *Proceedings of 16th International Conference on Pattern Recognition,* Vol.3, Iss., 2002 Pages: 469- 472 vol.3.

[79] Gandhi, T.; Kasturi, R.; Antani, S., "Application of Planar Motion Segmentation for Scene Text Extraction," *Proceedings of 15th International Conference on Pattern Recognition,* Vol.1, Iss., 2000 Pages:445-449 vol.1.

[80] Myers; G. K., Bolles; R. C., Luong; Q.-T., Herson; J. A., Aradhye; H. B., "Rectification and Recognition of Text in 3-D Scenes," *International*

*Journal on Document Analysis and Recognition, 2003*, pp. 147-158 vol. 7, 2005.

[81] Clark P. and Mirmehdi M., "Location and Recovery of Text on Oriented Surfaces," *In Proceedings of SPIE Conference on Document Recognition and Retrieval VII*, volume 3967, pages 267-277, January 2000.

[82] Clark P. and Mirmehdi M., "Estimating the Orientation and Recovery of Text Planes in a Single Image," *In Proceedings of the 12th British Machine Vision Conference*, pages 421-430. BMVA Press, September 2001.

[83] Zheng Zhang; Chew Lim Tan, "Restoration of Images Scanned from Thick Bound Documents", *Proceedings of 2001 International Conference on Image Processing,* Vol.1, Iss., 2001 Pages:1074-1077 vol.1.

[84] Brown, M.S.; Seales, W.B., "Document Restoration Using 3D Shape: A General Deskewing Algorithm for Arbitrarily Warped Documents," *Proceedings of Eighth IEEE International Conference on Computer Vision*, Vol.2, Iss., 2001 Pages:367-374 vol.2.

[85] Pollard; S., Pilu; M., "Building cameras for capturing documents," *International Journal on Document Analysis and Recognition*, 2003, pp. 123-137 vol. 7, 2005.

[86] Doncescu, A.; Bouju, A.; Quillet, V., "Former Books Digital Processing: Image Warping," *Proceedings of Workshop on Document Image Analysis,* Vol., Iss., 20 Jun 1997 Pages:5-9.

[87] Zheng Zhang; Chew Lim Tan, "Correcting Document Image Warping Based on Regression of Curved Text Lines," *Proceedings of Seventh International Conference on Document Analysis and Recognition,* Vol., Iss., 3-6 Aug. 2003 Pages: 589- 593 vol.1.

[88] Zhang, Z.; Lim, C.L.; Fan, L., "Estimation of 3D Shape of Warped Document Surface for Image Restoration," *Proceedings of the 17th International Conference on Pattern Recognition,* Vol.1, Iss., 23-26 Aug. 2004 Pages: 486- 489 Vol.1.

[89] Kanungo, T.; Haralick, R.M.; Phillips, I., "Global and Local Document Degradation Models," *Proceedings of the Second International Conference on Document Analysis and Recognition,* Vol., Iss., 20-22 Oct 1993 Pages:730-734.

[90] Ulges, A., Lampert, C. H., and Breuel, T. 2004. "Document Capture Using Stereo Vision," *In Proceedings of the 2004 ACM Symposium on Document Engineering,* ACM Press, New York, NY, 198-200.

[91] Zheng Zhang; Chew Lim Tan; Liying Fan, "Restoration of Curved Document Images through 3D Shape Modeling," *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Vol.1, Iss., 27 June-2 July 2004 Pages: I-10- I-15 Vol.1.

[92] Huiping Li; Doermann, D., "Superresolution-Based Enhancement Of Text In Digital Video," *Proceedings of 15th International Conference on Pattern Recognition,* Vol.1, Iss., 2000 Pages:847-850 vol.1.

[93] Capel D.; Zisserman, A., "Super-Resolution Enhancement of Text Image Sequences", *Proceedings of 15th International Conference on Pattern Recognition,* Vol.1, Iss., 2000 Pages:600-605 vol.1.

[94] Donaldson, K.; Myers, G.K., "Bayesian Super-Resolution Of Text In Video With A Text-Specific Bimodal Prior", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Vol.1, Iss., 20-25 June 2005 Pages: 1188- 1195 vol. 1.

[95] Baker, S.; Kanade, T., "Limits On Super-Resolution and How To Break Them", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.24, Iss.9, Sep 2002 Pages: 1167- 1183.

[96] Dalley, G.; Freeman, B.; Marks, J., "Single-Frame Text Super-Resolution: a Bayesian Approach", *International Conference on Image Processing*, Vol.5, Iss., 24-27 Oct. 2004 Pages: 3295- 3298 Vol. 5.

[97] Datong Chen; Shearer, K.; Bourlard, H., "Text Enhancement with Asymmetric Filter for Video OCR", *Proceedings of 11th International Conference on Image Analysis and Processing,* Vol., Iss., 26-28 Sep 2001 Pages:192-197.

[98] Haritaoglu, E.D.; Haritaoglu, I., "Real Time Image Enhancement and Segmentation for Sign/Text Detection", *Proceedings of 2003 International Conference on Image Processing,* Vol.3, Iss., 14-17 Sept. 2003 Pages: III-993-6 vol.2.

[99] Lucas, M. S.; Panaretos, A. et al.; "ICDAR 2003 Robust Reading Competitions: Entries, Results, And Future Directions," *International*

*Journal on Document Analysis and Recognition, 2003*, vol., no.pp. 105-122 vol. 7, 2005.

[100] Bishop, C. M.; *Pattern Recognition and Machine Learning*, Springer, 2006.

[101] Yoshimura, H.; Etoh, M.; Kondo, K.; Yokoya, N., "Gray-Scale Character Recognition by Gabor Jets Projection," *Proceedings of 15th International Conference on Pattern Recognition,* vol.2, no., pp.335-338 vol.2, 2000.

[102] Hartley; R. I. and Zisserman; A., *Multiple View Geometry in Computer Vision*, Cambridge University Press, UK, 2003.