Rochester Institute of Technology
RIT Digital Institutional Repository

Theses

2011

Page layout analysis and classification in complex scanned documents

Mustafa Erkilinc

Follow this and additional works at: https://repository.rit.edu/theses

Recommended Citation
Erkilinc, Mustafa, "Page layout analysis and classification in complex scanned documents" (2011). Thesis. Rochester Institute of Technology. Accessed from

This Thesis is brought to you for free and open access by the RIT Libraries. For more information, please contact repository@rit.edu.

PAGE LAYOUT ANALYSIS AND CLASSIFICATION FOR COMPLEX SCANNED DOCUMENTS

by

Mustafa Sezer Erkilinc

A thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

IN

ELECTRICAL ENGINEERING

Approved by:

Prof. _____

Thesis Advisor – Dr. Eli Saber

Prof. _____

Thesis Committee Member – Dr. Sohail Dianat

Prof. _____

Thesis Committee Member – Dr. Ferat Sahin

Prof. _____

Department Head – Dr. Sohail Dianat

Department of Electrical and Microelectronic Engineering
Kate Gleason College of Engineering
ROCHESTER INSTITUTE OF TECHNOLOGY
Rochester, New York
July 2011

Thesis Author Permission Statement

Title of Thesis:	***PAGE LAYOUT ANALYSIS AND CLASSIFICATION IN COMPLEX SCANNED DOCUMENTS***
Author:	**Mustafa Sezer Erkilinc**
Degree:	Master of Science
Program:	Electrical Engineering
College:	Kate Gleason College of Engineering

I understand that I must submit a print copy of my thesis or dissertation to the RIT Archives, per current RIT guidelines for the completion of my degree. I hereby grant to the Rochester Institute of Technology and its agents the non-exclusive license to archive and make accessible my thesis or dissertation in whole or in part in all forms of media in perpetuity. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Print Reproduction Permission Granted:

I, **Mustafa Sezer Erkilinc**, hereby **grant permission** to the Rochester Institute of Technology to reproduce my thesis or dissertation in whole or in part. Any reproduction will not be for commercial use or profit.

Author: _____ Date: _____

Inclusion in the RIT Digital Media Library Electronic Thesis & Dissertation (ETD) Archive:

I, **Mustafa Sezer Erkilinc**, additionally grant to the Rochester Institute of Technology Digital Media Library (RIT DML) the non-exclusive license to archive and provide electronic access to my thesis or dissertation in whole or in part in all forms of media in perpetuity.

I understand that my work, in addition to its bibliographic record and abstract, will be available to the world-wide community of scholars and researchers through the RIT DML. I retain all other ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation. I am aware that the Rochester Institute of Technology does not require registration of copyright for ETDs.

I hereby certify that, if appropriate, I have obtained and attached written permission statements from the owners of each third party copyrighted matter to be included in my thesis or dissertation. I certify that the version I submitted is the same as that approved by my committee.

Author: _____ Date: _____

DEDICATION

To my mother, *Şengül Ergünü*, for her endless support in my decisions.
To my father, *Gürsel Erkılınç*, for broadening my vision.
To my grandparents, *Mustafa & Huriye Ergünü*, for always being there for me.

DEDICATION

Anne ve Babama, ufkumu geniş tuttuklari ve verdikleri sonsuz destek için,
Anneanne ve Dedeme, her zaman yanmda olduklari için,
binlerce kez teşekkürler.

ACKNOWLEDGEMENTS

I am elated to have completed my graduate study. I have gained valuable knowledge in two years of graduate study at RIT. This experience has presented opportunities to improve myself both academically and personally. In this regard, I am grateful to RIT and the Department of Electrical and Microelectronic Engineering.

I would like to acknowledge and extend my heartfelt gratitude to the following colleagues: Mustafa Jaber for sharing his valuable experiences, and always being optimistic, Abdul Haleem Syed for his positive outlook on life and for our worthwhile discussions. I am also thankful to Mr. Peter Bauer, and the Hewlett Packard Company for their support and sponsorship for making this research possible.

I am very grateful to Dr. Drew Maywar who helped me discover a new world in electrical engineering. His knowledge and vision changed my point of view, and provided me with a new focus in an area of study which fascinates me. I also would like to thank to Dr. Ferat Sahin and Prof. Sohail Dianat for agreeing to be committee members for my thesis, and sharing their valuable ideas and reviews.

And finally, special thanks to Prof. Eli Saber who gave me an opportunity to work on his exciting projects. This thesis could not have been written without his encouragement throughout my academic life at RIT. His exacting character motivated me to aspire to greater academic achievements. He never accepted less than my best efforts, as he well knows. In addition to his academic support, he has always been there and willing to discuss my concerns, even if they are very personal. He taught me the significance of pushing myself, trying hard, and preserving.

Thank you Prof. Eli Saber for everything, it has been a pleasure and an honor being your student.

M. Sezer Erkilinc

ABSTRACT

Page layout analysis has been extensively studied since the 1980‘s, particularly after computers began to be used for document storage or database units. For efficient document storage and retrieval from a database, a paper document would be transformed into its electronic version. Algorithms and methodologies are used for document image analysis in order to segment a scanned document into different regions such as text, image or line regions. To contribute a novel approach in the field of page layout analysis and classification, this algorithm is developed for both RGB space and grey-scale scanned documents without requiring any specific document types, and scanning techniques. In this thesis, a page classification algorithm is proposed which mainly applies wavelet transform, Markov random field (MRF) and Hough transform to segment text, photo and strong-edge/line regions in both color and gray-scale scanned documents. The algorithm is developed to handle both simple and complex page layout structures and contents (text only vs. book cover that includes text, lines and/or photos). The methodology consists of five modules. In the first module, called pre-processing, image enhancements techniques such as image scaling, filtering, color space conversion or gamma correction are applied in order to reduce computation time and enhance the scanned document. The techniques, used to perform the classification, are employed on the one-fourth resolution input image in the CIEL*a*b* color space. In the second module, the text detection module uses wavelet analysis to generate a text-region candidate map which is enhanced by applying a Run Length Encoding (RLE) technique for verification purposes. The third module, photo detection, initially uses block-wise segmentation which is based on basis vector projection technique. Then, MRF with maximum *a-posteriori* (MAP) optimization framework is utilized to generate photo map. Next, Hough transform is applied to locate lines in the fourth module. Techniques for edge detection, edge linkages, and line-segment fitting are used to detect strong-edges in the module as well. After those three classification maps are obtained, in the last module a final page layout map is generated by using K-Means. Features are extracted to classify

the intersection regions and merge into one classification map with K-Means clustering. The proposed technique is tested on several hundred images and its performance is validated by utilizing Confusion Matrix (CM). It shows that the technique achieves an average of $\sim$85% classification accuracy rate in text, photo, and background regions on a variety of scanned documents like articles, magazines, business-cards, dictionaries or newsletters *etc.* More importantly, it performs independently from a scanning process and an input scanned document (RGB or gray-scale) with comparable classification quality.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Objectives and Motivations

In the late 1980's, document databases started to shift from hard-copy to soft-copy with the appearance of fast computers, large computer memories, and inexpensive scanners. They were stored digitally in large document databases and called document images. In the beginning of the 1990's, methodologies, algorithms and systems were invented and developed for document image analysis in order to extract information from document images in a "human-like" fashion. Extracting information from a document refers to locating and extracting line, photo or text regions hierarchically [1]. A hierarchy of document processing is illustrated in Fig. 1.1.

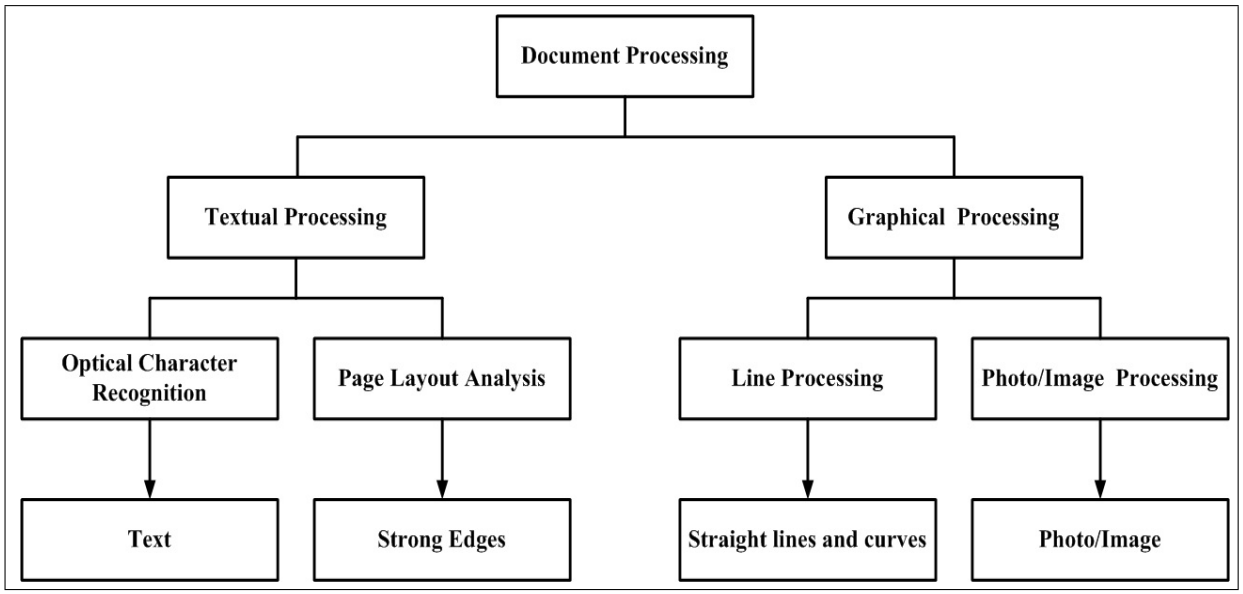


Fig. 1.1: A hierarchy of document processing. [1]

Today, the outcomes of research in document processing can be seen in many applications. Object-oriented rendering, extracting flowcharts and body diagrams from a scanned document for computer storage, document retrieval, query-images/texts and optical character recognition (OCR). Specifically, the millions of old paper volumes now in libraries will be replaced by computer

files in document images that can be searched for content. Signatures will be analyzed in the software-world for verification and security access [1]. These type of document analysis systems will enable to extract information without depending on the file formats. Automated mail-sorting and address recognition will become quicker and more accurate with text information extraction in the document. The number of mis-filed, mis-shelved or lost of material will be decreased by these document analysis techniques [2]. These examples serve as motivations for exploring potential solutions for document image analysis.

1.2 Literature Review

Document/Page segmentation is one of the topics researched in document processing to achieve homogeneity criteria for the connected regions of text, graphics and space. It is preferably used as an initial step for document structure analysis such as OCR [3, 4, 5] and document retrieval [6, 7, 8] [see also [9, 10] for comprehensive surveys in OCR and document retrieval].

There are three-main approaches in document segmentation. The top-down approach looks for global information on the page, such as black and white stripes, for the purpose of splitting the document image into blocks, blocks into lines, and lines into words. Fisher *et al.* presented the automatic segmenting of a document image which was enhanced by applying morphological operations, skew correction and adaptive filtering [11]. Then, the process continued with Run length encoding (RLE) algorithm to calculate the connected components' locations and statistics. Esposito *et al.* utilized the page layout feature which consists of geometrical characteristics in order to segment the image [12]. It was based on inductive generalization of a document layout style. Haralick *et al.* extended the scope of algorithm in [12] by adding various document images in the testing stage [13]. Automated text block extraction with image skew correction, which was based on a growing procedure guided by local information in complicated layout documents, was proposed by Zlatopolsky [14]. Sharma *et al.* also used a top-down approach by finding rectangular blocks in scanned documents and applying vertical and horizontal projections to a document

image [15]. Shi *et al.* proposed a top-down approach by using local connectivity property for document page segmentation [16].

The bottom-up approach starts with local information such as connected components in a specific region or block and first locates the words, then merges the words into lines, the lines into blocks and the blocks into columns. Wahl *et al.* utilized from run length algorithm to divide the page into rectangular regions [17]. Then, meaningful features are computed in these regions and a linear adaptive classification scheme is constructed to discriminate text regions from others. The Lam's *et al.* study employed a method which segments a newspaper document image into labeled macro zones and filters connected components to determine the content of the zones as text or non-text [18]. The drawback of [17] and [18] was that segmentation was achieved upon the assumption that the document image consists of rectangular areas. Antonacopoulos *et al.* proposed a technique that used the structure of the background white space, surrounded by the printed zones [19]. The benefit of the approach in [19] was that it did not make any assumptions about the shape or structure of the regions as opposed to [17] and [18]. It was capable of detecting complex shape regions more accurately than the existing methods. Drivas *et al.* incorporated a bottom-up document segmentation algorithm [20]. It utilized connected rectangular block based initial segmentation and then extracted simple histogram based features in order to determine textual and non-textual zones in a document image. Simon *et al.* developed and generalized the Kruskal's algorithm [21] and applied a special distance-metric between the components to construct a physical page structure [22]. The study reflected all the significant advantages of bottom-up systems such as being independent from text spacing and different block alignments. Jain *et al.* used traditional bottom-up approach based on the connected component extraction to achieve page segmentation and region identification [23]. Grover *et al.* extracted textual regions separating from the graphics portion by utilizing sharp edge features which were missing in image regions [24].

The hybrid approach achieves the document segmentation and classification based primarily on extracted features. The document image is subdivided into blocks and then required features

are computed. Jain *et al.* proposed a segmentation method for document images based on a multichannel filtering approach to texture segmentation [25]. Two-dimensional Gabor filters were used to extract texture features for text and non-text (image) regions. Duong *et al.* presented a document analysis system which segments the image as text and non-text zones [26]. The methodology retrieved a region of interest (ROI) from grey-scale document images via cumulative gradient considerations. Then, geometric and texture features are utilized in classification. In Randen and Husoy's study [27], a critically sampled filter bank was applied to the image, and local subband energy features were extracted to classify text/image regions by using K-Means algorithm. Fletcher *et al.* introduced a methodology which generated connected components [28]. Next, it grouped the connected components by using Hough transform into logical character strings in order to separate text from graphics. Tombre *et al.* extended the work in [28] to make it more applicable for graphics-rich documents by extracting features from an histogram of the connected components, filtering and thresholding [29]. Lin *et al.* utilized five Grey Level Co-occurrence Matrix (GLCM) that sub-divided the image into blocks to classify contents of document images as graphics, text and space [30]. Then, according to those features, connected blocks are clustered by applying K-Means.

Instead of selecting the features manually, they could be extracted by automatically which was presented in the work by Wang *et al.* [31]. This approach was an efficient and forward selection algorithm that iteratively constructed one linear feature at a time until a desired error rate was achieved. Although the proposed approach was applicable to many databases in literature, it was strongly data-driven and restricted to linear features. They improved their work in Wang's *et al.* study by extracting more features and evaluating on different document databases which contain images, graphics, handwriting and machine-printed text regions [32].

All the studies mentioned above solve the segmentation or classification problem in an unsupervised way. In other words, they did not require any *a-priori* information to achieve segmentation or classification. Another reason why unsupervised segmentation was employed particularly at the

end of the 80's and the beginning of 90's was the computation time issue. The training phase, not surprisingly, was very time consuming stage with 90's processor technology. However, this computation time issue began to be addressed with 2000's processor technology. Therefore, there was also research literature involving supervised segmentation or classification to solve the page segmentation problem. Chaudhury *et al.* presented a model-guided segmentation and document layout extraction scheme [33]. The proposed system extracted features which consist of contextual information and spatial configuration of a given document. It learned the relations between the layout specifications using Hierarchical Conditional Random Fields (CRFs). Baird *et al.* developed an automatically trainable methods for grey level and color document images which first obtained the pixel-wise features, trained them and then classified the regions by utilizing k-Nearest Neighbor (k-NN) learning technique [34]. Zheng *et al.* proposed a novel approach by treating noise as a separate class, modeling it based on selected features and classifying the text regions with trained Fisher classifiers [35]. Besides using Fisher classifiers, layout structure was obtained by using Markov Random Field (MRF) as a post-processing stage. Decision-tree classifiers and self-organizing maps were employed in the work of Shin *et al.* by using "visual similarity" of layout structure features such as content regions, column structures, relative point sizes of fonts *etc.* [36] Kumar *et al.* developed a novel approach for text segmentation in document images by applying globally matched wavelet filters [37]. The framework broadened to detect picture and background components in the image by combining multiple two-class Fisher classifiers and MRF formulation-based pixel labeling scheme to utilize from contextual information. In the core of the methodology, established by Caponetti *et al.*, neuro-fuzzy supervised technique was incorporated to perform the segmentation [38]. Initial segmentation was achieved by multi-scale processing, and a set of classical morphological operators was utilized for merging the pixels into coherent text, graphics or background regions. However, supervised techniques discussed above were computationally expensive when compared with unsupervised ones.

1.3 Contributions

In this thesis, a new unsupervised document classification algorithm is proposed. The benefits of the algorithm are the followings:

1) Robust classification for complex color and grey-scale scanned documents.

2) Independence from a type of scanned documents.

3) Utilization of both a textual map and a non-textual map to classify intersection regions of the text and photo map.

4) A potential solution that meets the computational efficiency constraint for most practical applications.

5) Independence from a scanning technique.

An overview of the proposed approach is shown in Fig. 1.2. The algorithm starts with a pre-processing module which applies image enhancements techniques such as image sizing, color space conversion, gamma correction and morphological operation. Sizing is applied to reduce the computation time and increase the speed to achieve the algorithm in real-time environment. Color-space conversion gamma correction are applied to color and grey-scale scanned documents,

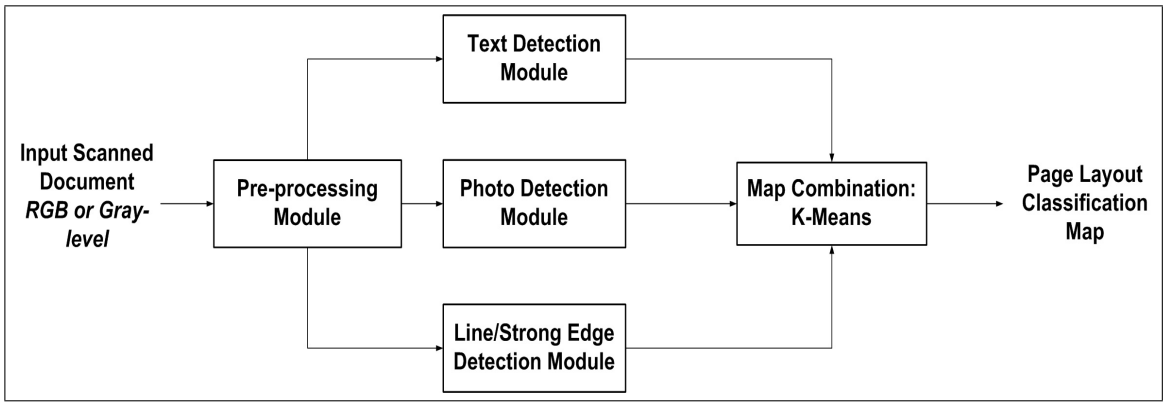


Fig. 1.2: Overview of the proposed approach.

respectively [see Fig. 1.2]. As a final stage in pre-processing module, to enhance the text components such as characters in text regions, dilation, called morphological operation, is performed on lightness (L*) component of CIEL*a*b* color space, called enhanced L* channel. All the steps in the pre-processing module mentioned above, are explained explicitly in Section 3.1. Three different modules follow the pre-processing stage to detect text, photo/image and strong edge/line regions. The modules process the enhanced L* component of the image separately. The main core of the photo detection stage relies on block-wise segmentation, proposed by Won [39], and MRF - maximum *a-posteriori* (MAP) optimization segmentation. Wavelet decomposition [40] is utilized to extract features and classify text regions in the document image. For confirmation purpose, Run-Length Encoding (RLE) is applied to obtain a final text map. Strong edge/line regions are extracted by employing the Hough transform on the enhanced L* channel. The resultant three separate classification maps are combined with the K-Means clustering to classify the blocks/pixels either text, photo or strong edge/line and merge into one final map. Although it converges to a local solution, it has low computationally complexity and provides satisfying result. The fundamental concept and detailed explanation of the modules can be found in Background [see Chapter 2] and Proposed Algorithm [see Chapter 3] chapters.

The aforementioned procedure is applied to both simple and complex background document images. In addition, it provides a solution for both color and grey-scale scanned documents. The proposed algorithm is entirely implemented in MATLAB® and tested on a large database of $\sim$700 document images which are scanned with *300* dots per inch (dpi). The performance evaluations show that the proposed algorithm is robust and accurate enough for the applications discussed in Section 1.4, and less computationally complexity.

The work discussed in this thesis is published in the following organizations.

1) M. S. Erkilinc, M. Jaber, E. Saber, and P. Bauer, "Page layout analysis for complex scanned documents", SPIE Opt. Eng. + Apps. Conf.: Apps. of Digital Image Proc. XXXIV, 2011.

2) M. S. Erkilinc, E. Saber and P. Bauer, "Page layout analysis for complex scanned documents",

SPIE Newsroom, Electronic Imaging and Signal Processing.

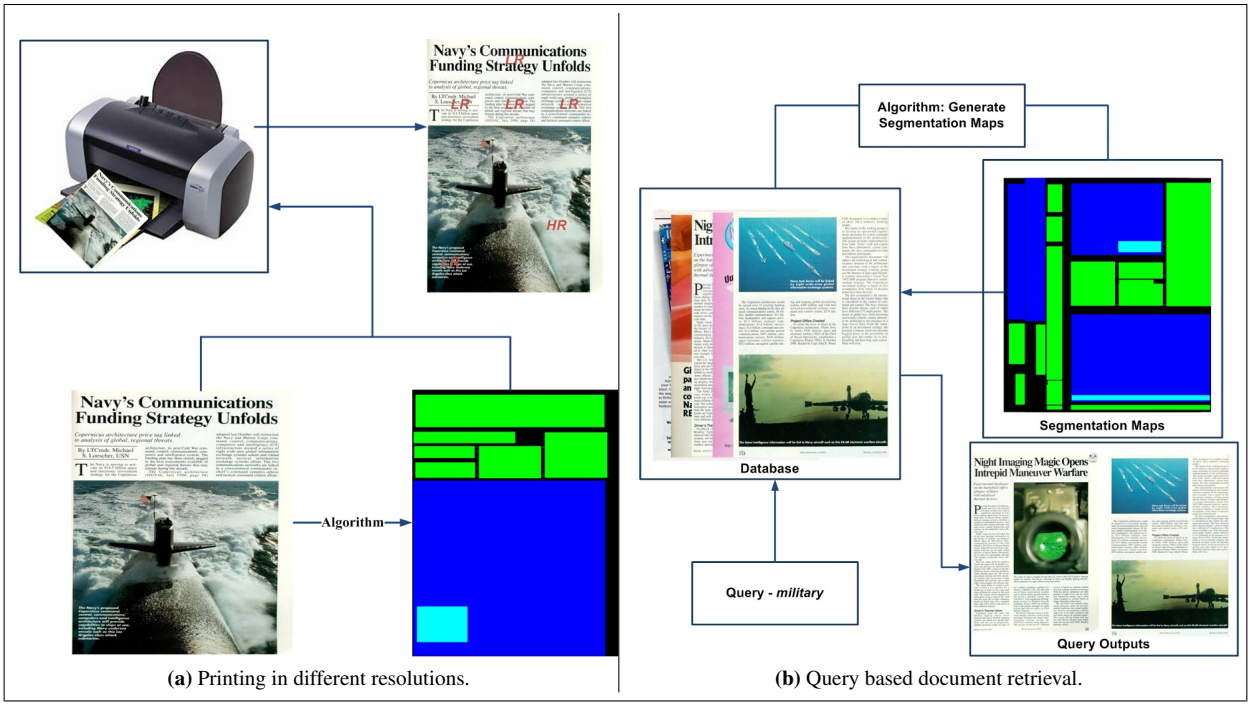
1.4 Potential Applications

Document segmentation has been widely used in OCR, document retrieval process where an efficient memory consumption and a quick retrieval are required. In addition, database update and efficient cartridge usage while printing the documents in different resolutions are other typical application. The proposed methodology is developed for a commercial purpose emphasizing performance time and accuracy. In this section, a few applications are presented which could utilize the proposed algorithm.

1.4.1 Content based document retrieval

One of the objectives of document classification application, and of document image analysis in general, is to recognize and extract text and graphic components for use by people throughout the world. Today, imaging systems, particularly scanners, are used to store great numbers of document images in databases so they can be retrieved. Additionally, different resolutions can be embedded while printing the documents by extracting text and graphic regions. This provides a better and more efficient print quality. A typical document retrieval framework and a methodology for better print quality, are demonstrated in Fig. 1.3.

The scanned document image printed is classified by using the proposed algorithm. Then, both the classification map and the input image are sent to the printer to print the document image in different resolutions. The proposed technique detects text, photo and strong egde/line regions. While text regions are printed in low resolution (LR), photo and strong edge/line regions are printed in high resolution (HR). The demonstration for the system is given in Fig. 1.3(a). There are five text regions which cover almost half of the document, and one image which covers the rest of it [see Fig. 1.3(a)]. By using low resolution to print the textual areas, a considerable amount of ink can be saved.



(a) Printing in different resolutions. **(b)** Query based document retrieval.

Fig. 1.3: Content based document processing applications.

From a user perspective, high resolution is not required for text regions, because a reader or a user does not need a high resolution to read them. The only requirement is that they are legible. Since low resolution can meet this requirement, printing the regions in high resolution wastes ink. On the other hand, photo and strong-edge/line regions should be printed in high resolution in order to satisfy the user in terms of quality. The aim is to use printers more efficiently and prevent cartridge or laser waste.

Furthermore, the same approach can be utilized for document retrieval, illustrated in Fig. 1.3(b) in large databases. After classification is achieved, extracted components can be labeled with a group name which best describes them. This yields to group the documents under the same group name as well. The labeling process provides faster access to the documents in the database. In other words, the several megabytes (MB) of raw data in the database can be culled in a much more concise way by assigning semantic or logical description to the extracted components (textual and graphical regions).

Depending on the query, requested documents can be retrieved by extracting their functional

parts while the database stores the original document images with their classification maps. For instance, over a million pieces of mail are handled in one day by the United States post office. The proposed document classification method to perform sorting mail according to its image content and/or address recognition (text content) would be especially useful for processing this volume of mail more quickly, and accurately.

1.4.2 Optical character recognition

Two primary categories, presented in Fig. 1.1 sum up the entire field of document processing: Textual processing deals with text zones (particularly achieving OCR) and finding columns, paragraphs, text lines, and words. Besides this, graphical processing deals with images, photos, block diagrams, tables and logos, *etc.* OCR is an electronic translation of scanned images of handwritten, typewritten or printed text into machine-encoded text. OCR techniques, presented in Fig. 1.4 below, are widely used in detecting and examining signatures at banks, converting books and documents into electronic files, or publishing text on a web-site.

Three main modules, discussed in Chapter 3, generate three different classification maps. For OCR, text map can be broken into paragraphs, words and text characters depending on the query. For instance, the word "MELIOS" is considered as a query, and it is a military term. The hierarchal order of how it is recognized is exhibited Fig. 1.4. The query can be anything as long as it can be represented in a textual region. The framework is provided to edit the text, search for a word or phrase, store it more compactly, and apply techniques such as machine translation, text-to-speech and text mining to it.

1.5 Thesis Outline

The remainder of the thesis is organized as follows: Chapter 2, Background, a review of the concepts is outlined and examined to implement the proposed algorithm successfully. The proposed

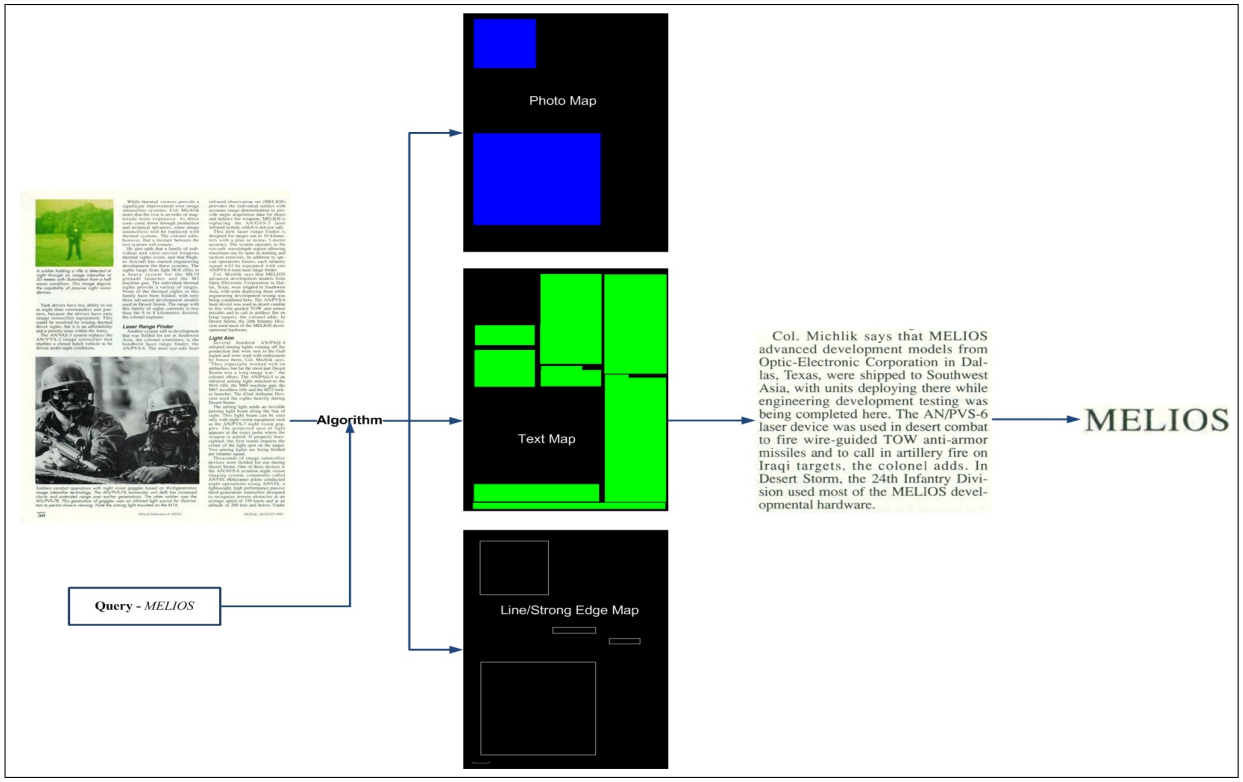


Fig. 1.4: Optical character recognition system.

methodology is introduced in Chapter 3 and consists of five sections: Pre-processing, Text Detection, Photo Detection, Line Detection and Map Combination. Section 3.1 explains the preprocessing module that is performed before the main core of the algorithm to enhance the image, increase accuracy of the algorithm and obtain lower computation time. Section 3.2, 3.3,and 3.4 describe how text, photo and strong edge/line map are obtained. Combination and merging procedure of the maps are described in Section 3.5. Experimental results are presented and performance of the segmentation algorithm is evaluated by confusion matrix methodology in Chapter 4. Conclusions are drawn in Chapter 5.

CHAPTER 2: BACKGROUND

This chapter clarifies some technical concepts, and provides technical background for the algorithms that are employed in the proposed document classification technique. Wavelet transform (WT), MRF, Hough transform and RLE algorithms, which are the main parts of text, photo and strong edge/line detection modules, are discussed with their mathematical insight. First, wavelet transform and MRF are discussed separately because of the individual roles they play in the art of image processing. Moreover, the formulation of Hough transform and RLE algorithms, and how they are applied to the images, is explained.

2.1 Wavelet Transform

In mathematics, a wavelet series is a representation of a square-integrable (real-or complex-valued) function by certain orthonormal series generated by a wavelet. It is a powerful tool which can decompose a signal into various frequency bands. These bands are generally taken as horizontal, vertical, and diagonal spatial frequency characteristics of the data. Basics of the wavelet transform are presented in Section 2.1.1 to establish the foundation for multiresolution image representation. Furthermore, discrete conversion of the theory and its performance in discrete domain are explained explicitly. Detailed mathematical analysis of initial multiresolution image representation both in 1-D and 2-D, introduced by Mallat [41], is given in Section 2.1.2.

2.1.1 Fundamentals of wavelet transform

Every vector in a vector space can be represented as a linear combination of the basis vectors in that vector space. This concept can be easily generalized to functions by replacing the basis vectors with basis functions.

$$f(t) = \sum_{k} \mu_k \phi_t(k), \tag{2.1}$$

where $f(t)$, and $\phi_t(k)$ the signal and basis function, respectively. μ_k is the corresponding basis function coefficient. Furthermore, basis functions in any domain should both be orthogonal to span the space completely and orthonormal for normalization purpose, as given in Eq. 2.2. Additionally, it has sufficient properties for reconstruction, recovering the signal by inverse wavelet transform.

$$\begin{aligned} < f(t), g(t) > &= \int_a^b f(t)g^*(t)dt = 0, \\ < f(t), f(t) > &= \int_a^b f(t)f^*(t)dt = 1. \end{aligned} \tag{2.2}$$

where $*$ denotes complex conjugation. According to the above definition of the inner product, the Continuous WT(CWT) can be written as the inner product of the signal and basis function(wavelet) in Eq. 2.3.

$$CWT_f^{\Psi}(\tau, s) = \Gamma_f^{\Psi}(\tau, s) = \int_a^b f(t)\Psi_{\tau,s}^*(t)dt, \tag{2.3}$$

where $\Gamma_f^{\Psi}(\tau, s)$ and $\Psi(t)$ are the wavelet coefficient and mother wavelet, respectively. Mother wavelet, defined in Eq. 2.4 , is the basic wavelet which is utilized to generate different wavelets by scaling and translation operations [42].

$$\Psi_{\tau,s} = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-\tau}{s}\right), \tag{2.4}$$

where s and τ are the scale and translation factors. The factor of $s^{-1/2}$ stands for the energy normalization. Eq. 2.3 shows how a signal, $f(t)$, is decomposed into a set of wavelets, $\Psi_{\tau,s}$. The variables s and τ are the new dimensions, scale and translation, after the wavelet transform. The expression for the CWT, given in Eq. 2.3, shows that the wavelet analysis is a measure of similar frequency content between the wavelets, $\Psi_{\tau,s}$, and the signal, $f(t)$. CWT coefficients, $\Gamma_f^{\Psi}(\tau, s)$, represent the degree of closeness of a signal to a wavelet. Thus, the inner product of a wavelet and signal will give a relatively large number, corresponded to the $\Gamma_f^{\Psi}(\tau, s)$ if the signal is represented better by the employed wavelet in the inner product operation in Eq. 2.3. As previously mentioned before, original signal can be also reconstructed by integrating the wavelet coefficients

and corresponding wavelets in the domain of S and T, shown in Eq. 2.5 since WT is an invertible operation [42].

$$f(t) = \int_{s \in S} \int_{\tau \in T} \Gamma_x^{\Psi}(\tau, s) \Psi_{\tau,s}^*(t) d\tau dt. \tag{2.5}$$

To enable the reconstruction without loss of information, square integrable functions (wavelets), $\Psi(t)$, should satisfy the *admissibility condition*, given in Eq. 2.6.

$$F\left[\Psi(\omega)\right] = \int_{-\pi}^{\pi} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega =< +\infty. \tag{2.6}$$

In Eq. 2.6, $F\left[\Psi(\omega)\right]$ denotes for the Fourier transform (FT) of a wavelet. The admissibility condition implies that the FT of $\Psi(t)$ vanishes at the zero frequencies, *i.e.*

$$|\Psi(\omega)| \Big|_{\omega=0} = 0. \tag{2.7}$$

In other words, wavelets must have have a band-pass like spectrum. Eq. 2.7 also indicates that the average value of the wavelet in the time domain must be zero,

$$\int_{-\infty}^{\infty} \Psi(t) dt = 0. \tag{2.8}$$

Therefore it must be oscillatory or a *wave*. As shown in Eq. 2.3, the 1-D WT is fundamentally 2-D because of the parameters, τ and s. Thus the 2-D WT is four-dimensional. The time-bandwidth product of the WT increases exponentially since the input signal is squared. Hence, a scale, denoted by s, is added to ensure that the WT is decreasing rapidly [42].

Moreover, the wavelet function should have some smoothness and concentration in both time and frequency domains. Regularity condition is explained by using the concept of vanishing moments. The wavelet transform, given in Eq. 2.3, can be defined by utilizing the Taylor series at $t = 0$ until the order n and letting $\tau = 0$ for simplicity [43].

$$\Gamma(s,0) = \frac{1}{\sqrt{s}}\left[\sum_{m=0} nf^{(m)}(0)\int \frac{t^m}{m!}\Psi\left(\frac{t}{s}\right)dt + O(n+1)\right], \tag{2.9}$$

where $f^{(m)}$ is the m^{th} derivative of f and $O(n+1)$ stands for other orders in expansion. Also, moments of a wavelet M_m is defined in Eq. 2.10 as,

$$M_m = \int t^m \Psi(t)dt. \tag{2.10}$$

So, Eq. 2.9 can be re-written by using Eq. 2.10,

$$\Gamma(s,0) = \frac{q}{\sqrt{s}}\left[f(0)M_0 s^1 + \frac{f^{(1)}(0)}{1!}M_1 s^2 + \frac{f^{(2)}(0)}{2!}M_2 s^3 + \cdots + \frac{f^{(n)}(0)}{n!}M_n s^{n+1} + O(s^{n+2})\right]. \tag{2.11}$$

From the admissibility condition, it is already known that the 0^{th} moment, $M_0 = 0$, so that Eq. 2.11 is simplified as shown in Eq. 2.12,

$$\Gamma(s,0) = \frac{q}{\sqrt{s}}\left[\frac{f^{(1)}(0)}{1!}M_1 s^2 + \frac{f^{(2)}(0)}{2!}M_2 s^3 + \cdots + \frac{f^{(n)}(0)}{n!}M_n s^{n+1} + O(s^{n+2})\right]. \tag{2.12}$$

If a wavelet has N vanishing moments, then the approximation order of the wavelet transform is also N since the moments up to M_n are very small values compared to N. Then, the wavelet transform coefficients, $\Gamma(s,\tau)$, can diminish as fast as s^{n+2} for a smooth signal [44].

So far, the discussion has dealt with only continuous case. It is essential to convert it into discrete case for practical purposes. Unlike continuous wavelets, discrete wavelets can only be scaled and translated in discrete steps. This can be achieved by modifying the mother wavelet, given in Eq. 2.4.

$$\Psi_{æ,k}(t) = \frac{1}{\sqrt{s_o^j}}\Psi\left(\frac{t - k\tau_0 s_0^j}{s_0^j}\right), \tag{2.13}$$

where j and k are integers and $s_0 > 1$ is a fixed dilation step. The translation factor, τ_0, depends

on the dilation step. The effect of discretizing the wavelet is that the time domain is now sampled at discrete intervals. s_0 is usually chosen as 2 in order to correspond to the dyadic sampling for the sampling of a frequency axis. In the same manner, the translation factor, τ_0, is assumed to be 1 to have dyadic sampling of the time axis as well. These assumptions are reasonable for practical applications involving computers [43].

A continuous signal can be represented as a series of discrete wavelet coefficients, called wavelet series decomposition, but this decomposition is must be reversible. In other words, the continuous signal can be reconstructed by its own wavelet series decomposed signals using its own discrete wavelet coefficients. For stable reconstruction, the energy of the wavelet coefficients should be bounded by two positive numbers, shown in Eq. 2.14.

$$A \left\| f \right\|^2 \leq \sum_{j,k} \left| \langle f, \Psi_{j,k} \rangle \right|^2 \leq B \left\| f \right\|^2, \tag{2.14}$$

where $\left\| f \right\|^2$ is the energy of $f(t)$, $A > 0\, B < \infty$ while A, B are independent of $f(t)$. In addition to stability condition, orthogonality and orthonormality conditions should be satisfied for reconstruction. As previously discussed, discrete wavelets can be represented in terms of mother wavelets so that they can satisfy the condition given in Eq. 2.2 by adjusting scaling and translation constants in mother wavelets [43].

$$\int \Psi_{j,k}(t) \Psi^*_{m,n}(t) dt = \begin{cases} 1 & \text{if } j = m \quad \text{and} \quad k = n \\ 0 & otherwise \end{cases}. \tag{2.15}$$

If all the conditions mentioned above are satisfied, any type of signal can be reconstructed by the linear combination of wavelet basis functions weighted by the wavelet coefficients.

Still, the process needs an infinite number of scaling and translation constants to calculate the wavelet transform since the signal spectrum is infinite. Wavelets have a band-pass like spectrum [see Eq. 2.7] which serves as a finite spectrum. Hence, an infinite set of wavelets with infinite spectrum can be replaced by finite scaling function. This property yields a conclusion that if

one wavelet can be seen as a band-pass filter, then a series of dilated wavelets can be seen as a band-pass filter bank, which covers all spectra of the signal as well. The filter bank covers all spectrum according to the center frequencies and the width of each filter spectrum depends on a ratio, referred as the *fidelity factor*, Q. Eventually, if the wavelet transform is assumed to be a filter bank, then taking a wavelet transform of a signal can be considered as passing the signal through this filter bank, called *sub-band coding*. The output of each filter stage in the filter bank gives the wavelet and scaling function transform coefficients. The technique discussed above is illustrated in Fig. 2.1 [41].

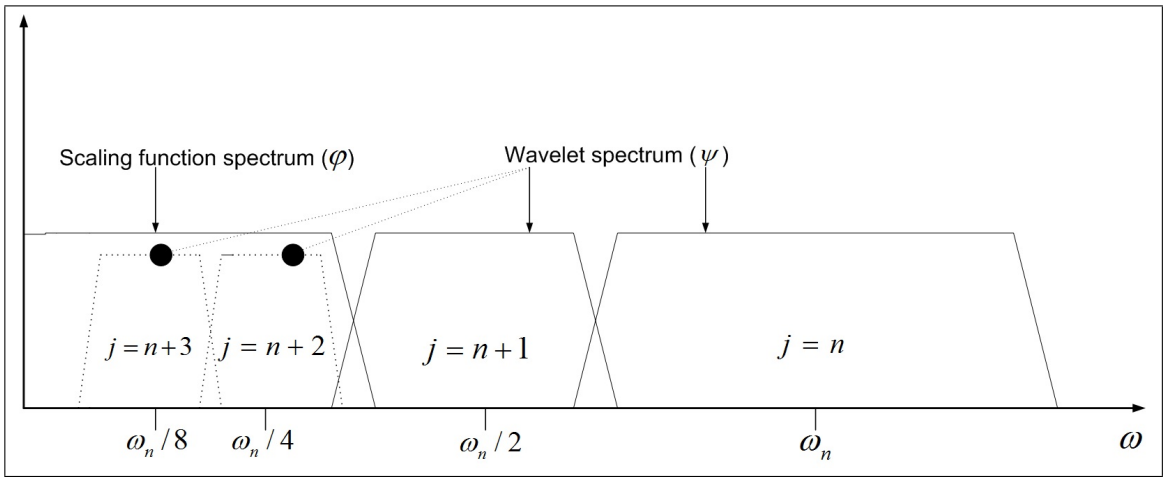
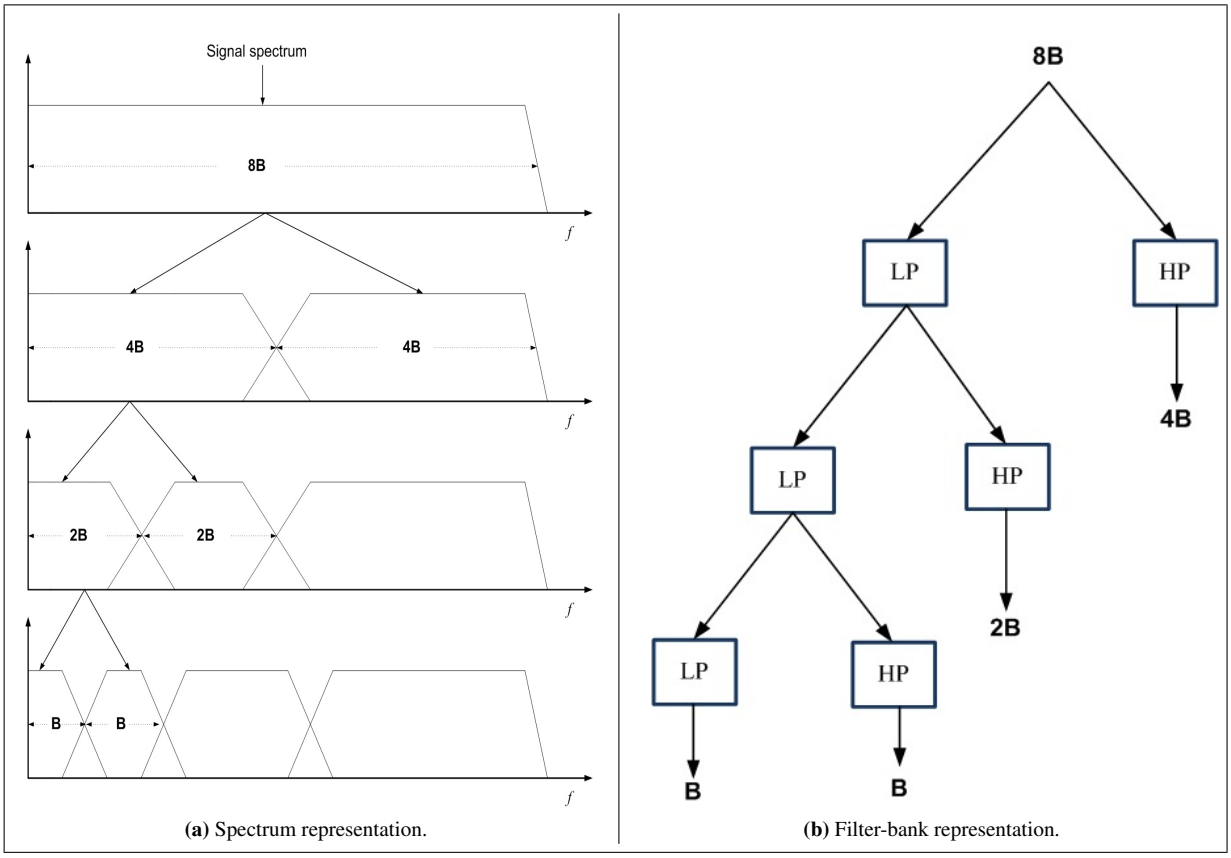


Fig. 2.1: Replacing one scaling function instead of infinite a set of wavelets [41].

The sub-band coding scheme can be implemented in two ways. One approach is to build many band-pass filters independently. Although it provides freedom in determining the width of each band, all the filters should be designed separately to segment the spectrum into different frequency bands. Thus, it requires extensive computation time. Another approach is to split the entire spectrum of the signal hierarchically into two equal parts, a low-pass and a high-pass part at each filter stage. While the low-pass part is covering relatively flat or smooth surfaces, the high-pass part covers details such as edges or transition regions. As a continuation of the procedure, if the low-pass part at a certain filter stage contains some details, not captured in the high pass part, the signal can continue splitting in two equal parts iteratively to obtain more features (detail information) from

the signal or image. This explained scheme requires a design of only two filters. However, the filter spectrum width is fixed in the process, unlike the previous coding scheme. The demonstration for 1-D signal is depicted in Fig. 2.2. In Fig. 2.2(a), splitting a 1-D signal spectrum into various bands is represented and the corresponding filter operation is shown in Fig. 2.2(b).



(a) Spectrum representation. **(b)** Filter-bank representation.

Fig. 2.2: Splitting 1-D signal spectrum with an iterated filter bank [42].

In the previous paragraph, it is assumed that taking a wavelet transform of a signal is achieved passing the signal through a filter bank. In this manner, while the wavelets provides the band-pass bands, the scaling functions resultantly represent the low-pass band. Thus, a wavelet transform is a sub-band coding scheme using a constant-Q filter bank. This decomposition technique is called multiresolution analysis, explained explicitly in Section 2.1.2, since it decomposes a signal in different resolutions [41].

2.1.2 Multiresolution analysis

A set of wavelets can be estimated by employing a scaling function to cover the entire wavelet spectra in discrete case as it is depicted in the previous section. In this purpose, if the scaling function is treated as a signal, it can be decomposed into its wavelet components like in Eq. 2.16:

$$\psi(t) = \sum_{j,k} \Gamma(j,k)\Psi_{j,k}(t). \tag{2.16}$$

However, Eq. 2.16 still uses an infinite number of wavelets up to a certain scale j. If a wavelet spectrum is added to the scaling function spectrum, a new scaling function with a spectrum twice as wide as the first is obtained. This expression is formulated in Eq. 2.17. The effect of this addition can be expressed in terms of the first scaling function, shown in Eq. 2.16 by summing the second scaling function with the first scaling function [41].

$$\psi(2^j t) = \sum_{k} h_{j+1}(k)\psi(2^{j+1}t - k). \tag{2.17}$$

In Eq. 2.17, a set of wavelets can be also expressed in terms of the first scaling function so that a set of wavelets in each decomposition level can be replaced by a translated scaling functions. The wavelet at level j can be written as;

$$\Psi(2^j t) = \sum_{k} g_{j+1}(k)\psi(2^{j+1}t - k), \tag{2.18}$$

where $\Psi(2^j t)$ and $g_{j+1}(k)$ are the wavelet and band-pass filter. Eq. 2.5 implies that a signal, $f(t)$, can be expressed in terms of dilated (scaled) and translated wavelets up to a level $j-1$. This yields the result that $f(t)$ can be also expressed in terms of dilated and translated scaling function at a level j.

$$\Psi(2^j t) = \sum_{k} g_{j+1}(k)\psi(2^{j+1}t - k) + \sum_{k} \Gamma_{j-1}(k)\Psi(2^{j-1}t - k), \tag{2.19}$$

where the first term, $\sum_k g_{j+1}(k)\psi(2^{j+1}t - k)$, represents the signal in terms of scaling functions

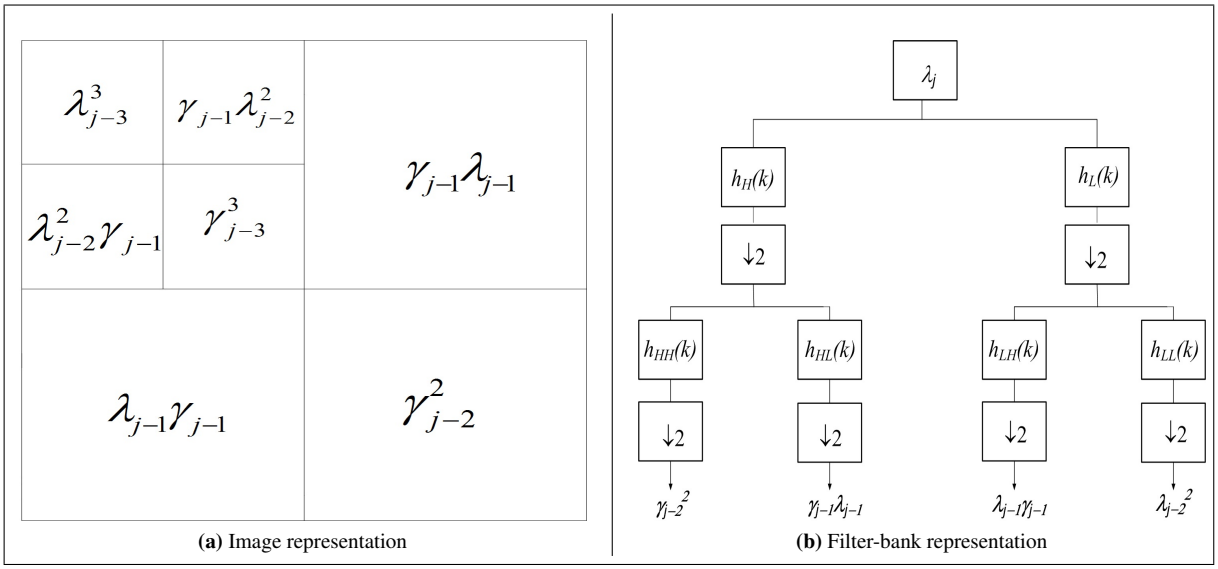
up to a level $j-1$ and the second term, $\sum_k \Gamma_{j-1}(k)\Psi(2^{j-1}t-k)$, denotes the signal at level j in terms of wavelets. If the scaling functions, $\psi_{j,k}(t)$ and $\Psi_{j,k}(t)$, are orthonormal to each other, then the coefficients, $\lambda_{j-1}(k)$ and $\Gamma_{j-1}(k)$, can be found by using Eq. 2.20.

$$\begin{aligned} \lambda_{j-1}(k) &= \langle f(t), \psi_{j,k}(t) \rangle \\ \Gamma_{j-1}(k) &= \langle f(t), \Psi_{j,k}(t) \rangle \end{aligned} \tag{2.20}$$

where the wavelet, $\Gamma_{j-1}(k)$, and scaling function, $\lambda_{j-1}(k)$, coefficients are expressed in a closed-form solution. By combining Eq. 2.17 and Eq. 2.18 with Eq. 2.20, open-form solution can be written as;

$$\begin{aligned} \lambda_{j-1}(k) &= \textstyle\sum_m h(m-2k)\lambda_j(m) \\ \Gamma_{j-1}(k) &= \textstyle\sum_m g(m-2k)\Gamma_j(m) \end{aligned}. \tag{2.21}$$

These two equations indicate that the wavelet and scaling function coefficients on a certain scale can be found by calculating a weighted sum of the scaling function coefficients from the previous scale. Eq. 2.21 means that the weighting factors of $h(k)$ corresponds to a low-pass filter since the coefficients, $\lambda_j(k)$, originates from the low-pass part of the splitted signal spectrum.



(a) Image representation **(b)** Filter-bank representation

Fig. 2.3: Multiresolution image representation [43].

The weighting factors of $g(k)$ corresponds a high-pass filter since the coefficients, $\Gamma_j(k)$, originates from the high-pass part of the splitted signal spectrum. From an digital signal processing perspective, one stage of an *iterated digital filter bank* can be established by using $h(k)$ and $g(k)$. In Fig. 2.3 above, implementation of the iterated digital filter bank is demonstrated by using 2-D signal (image). In the figure, $h(m-2k)$ and $g(m-2k)$ in Eq. 2.21 are denoted as $h_L(k)$ and $h_H(k)$, respectively.

2.2 Markov Random Field Modeling

A random field is a 2-D sample sequence where each sample is a random variable. Each outcome in the sample space produces a realization of the random field. MRF theory provides a useful and consistent modeling for context dependent units such as image pixels. It characterizes the units depending on the effect of neighborhood units using conditional MRF distributions. This theory was established by Hammersley and Clifford [45] in 1971, and developed by Besag [46].

The joint distribution information is required for MRF modeling in most applications. However, deriving the joint distribution from conditional distributions is a very difficult problem for MRFs. Hence, Gibbs distribution (GD) is incorporated with MRFs, called MRF-GD theorem, to obtain the joint distribution from conditional distributions. The theorem provides a mathematical convenience in terms of statistical signal processing in applications such as image analysis [47].

The MRF theory is utilized to model a *a-priori* probability of context dependent patterns. A particular MRF model favors its own labeled class of patterns by assigning larger probabilities than other pattern classes. Maximum *a-posteriori* (MAP) probability is one of the most popular statistical criteria for optimality. The MRF-MAP framework, introduced by Geman and Geman [48], develops algorithms to solve various vision problems such as image and video processing using probabilistic approach [49].

The objective is to maximize the joint posterior probability of the MRF labels in MAP-MRF

framework. The framework is constituted and the parameters are selected by using *Bayes* formulation in which the objective function depends on the joint prior distribution of the labels, the conditional probability of the observed data, and the probability of the observed data.

Section 2.2.1 outlines the fundamentals of *Bayes* estimation. The MRF-MAP modeling process of the conditional probability of the observed data is discussed in Section 2.2.2, and Section 2.2.3 explains the joint prior distribution of the labels explicitly.

2.2.1 Bayes estimation

Bayes theorem is a fundamental theory in estimation theory. It points out that when both prior distribution and the likelihood function of a pattern are known, the best estimation can be obtained by Bayes labeling. The MAP optimization is a special case of the Bayes theorem. In Bayes estimation, a risk is minimized in order to obtain the optimal estimate. The objective function, risk, is defined as:

$$O(\hat{x}) = \int_{x \varepsilon F} C(\hat{x}, x) P(x|I) dx, \tag{2.22}$$

where I is the data, $C(\hat{x}, x)$ is a cost function, $P(x|I)$ is the conditional posterior distribution and F is the 2-D field. By using the Bayes rule, $P(x|I)$ can be calculated by using Eq. 2.23.

$$P(x|I) = \frac{p(I|x)P(x)}{p(I)}, \tag{2.23}$$

where $P(x|I)$ is *a-priori* probability of labeling of I, also referred as segmentation map in image processing, $p(I|x)$ is the *a-posteriori* probability or likelihood function of x for fixed I and $p(I)$ is the *a-priori* probability of I, which is given and can be assumed as deterministic in the problem. The cost function, $C(\hat{x}, x)$, is the error constraint while determining how the estimation is accurate for actual data, x. There are two popular choices for the cost function. The one which is based on the Euclidean distance, shown below:

$$C(\hat{x}, x) = \|\hat{x} - x\|^2 . \tag{2.24}$$

Second one, called $\epsilon(0-1)$, is given in Eq. 2.25,

$$C(\hat{x}, x) = \begin{cases} 0 & \text{if } \|\hat{x} - x\| \leq \epsilon \\ 1 & otherwise \end{cases} \tag{2.25}$$

where $\epsilon > 0$ is any small constant. If Eq. 2.24 is placed in Eq. 2.22, the variance of the estimate can be writen as follows;

$$O(\hat{x}) = \int_{x \varepsilon F} \|\hat{x} - x\|^2 P(x|I) dx. \tag{2.26}$$

The minimal variance estimate which is the mean of the posterior probability can be found by letting $\frac{\delta O(\hat{x})}{\delta \hat{x}} = 0$,

$$\hat{x} = \int_{x \varepsilon F} x P(x|I) dx. \tag{2.27}$$

For the ϵ cost function, the objective function becomes

$$O(\hat{x}) = 1 - \int_{x:\|\hat{x}-x\| \leq \epsilon} P(x|d) dx, \tag{2.28}$$

when $\epsilon \to 0$. Furthermore, Eq. 2.28 can be approximated by

$$O(\hat{x}) = 1 - \zeta P(x|d). \tag{2.29}$$

where ζ is the volume of the space containing all points x for which $\|\hat{x} - x\| \leq \epsilon$. Minimizing the cost function, given in Eq. 2.29, corresponds to a maximization of $P(x|I)$;

$$\hat{x} = arg \max_{x \in F} P(x|I), \tag{2.30}$$

which is known as the MAP estimate. $P(x|I)$ is proportional to the joint distribution, shown in Eq. 2.31 since $p(I)$ in Eq. 2.23 refers to an image whose probability is 1,

$$P(x|I) \propto P(x, I) = p(I|x)P(x). \tag{2.31}$$

Then, the MAP estimate is equivalently found as follows:

$$\hat{x} = arg\,\underset{x \in F}{max}\,[\,p(I|x)P(x)\,]\,. \tag{2.32}$$

2.2.2 Modeling conditional probability distribution (likelihood) function

As mentioned in Bayes estimation section, $P(x|I)$ is the posterior distribution of a MRF. The derivation of the distribution can start with a simple assumption. Let us assume that image surfaces are flat, then joint prior distribution of x can be written as follows;

$$P(x) = \frac{1}{Q} e^{-E(x)}, \tag{2.33}$$

where $E(x) = \sum_i \sum_j (x_i - x_j)^2$ is the *prior energy* for a flat surface. The noise on this flat surface can be also assumed as Gaussian noise, $\omega_i = x_i + n_i$, where $n_i \sim N(\mu, \sigma^2)$ then the likelihood distribution can be written as shown in Eq. 2.34.

$$p(I|x) = \frac{1}{\prod_{i=1}^{M} \sqrt{2\pi\sigma^2}} e^{-E(I|x)}, \tag{2.34}$$

where

$$E(I|x) = \sum_{i=1}^{M} (\omega_i - x_i)^2 / 2\sigma^2 \tag{2.35}$$

is the *likelihood energy*. Since $Q = \prod_{i=1}^{M} \sqrt{2\pi\sigma^2}$ is the normalization factor, the posterior probability in an optimization process becomes,

$$P(x|I) \propto e^{-E(x|I)}, \tag{2.36}$$

where

$$E(x|I) = E(I|x) + E(I)$$
$$= \sum_{i=1}^{M}(x_i - \omega_i)^2/2\sigma_i^2 + \sum_{i=1}^{M}(x_i - x_{i-1})^2 \quad (2.37)$$

is the *posterior energy*. The MAP estimate is equivalently found by minimizing the posterior energy function,

$$\hat{x} = arg\,\min_{x} E(x|I). \quad (2.38)$$

2.2.3 Gibbs distribution

Gibbs Distribution (GD) is first introduced by Derin *et al.* [50], and utilized to model an image data. Since image is a discrete signal in computer applications, we are interested in discrete 2-D random fields. It is defined over a finite $N_1 x N_2$ rectangular lattice of points (pixels) which is also defined as $L = \{(i,j) : 1 \leq i \leq N_1, i \leq j \leq N_2\}$. As a second step, definition of a neighborhood system on lattice, L, and the associated cliques are presented below.

Definition1: A collection of subsets of L can be described as;

$$\eta = \{\eta_{i,j} : (i,j) \in L, \eta_0 \subseteq L\} \quad (2.39)$$

which is a neighborhood system on L if only $\eta_{i,j}$ the neighborhood of pixel(i,j) is such that

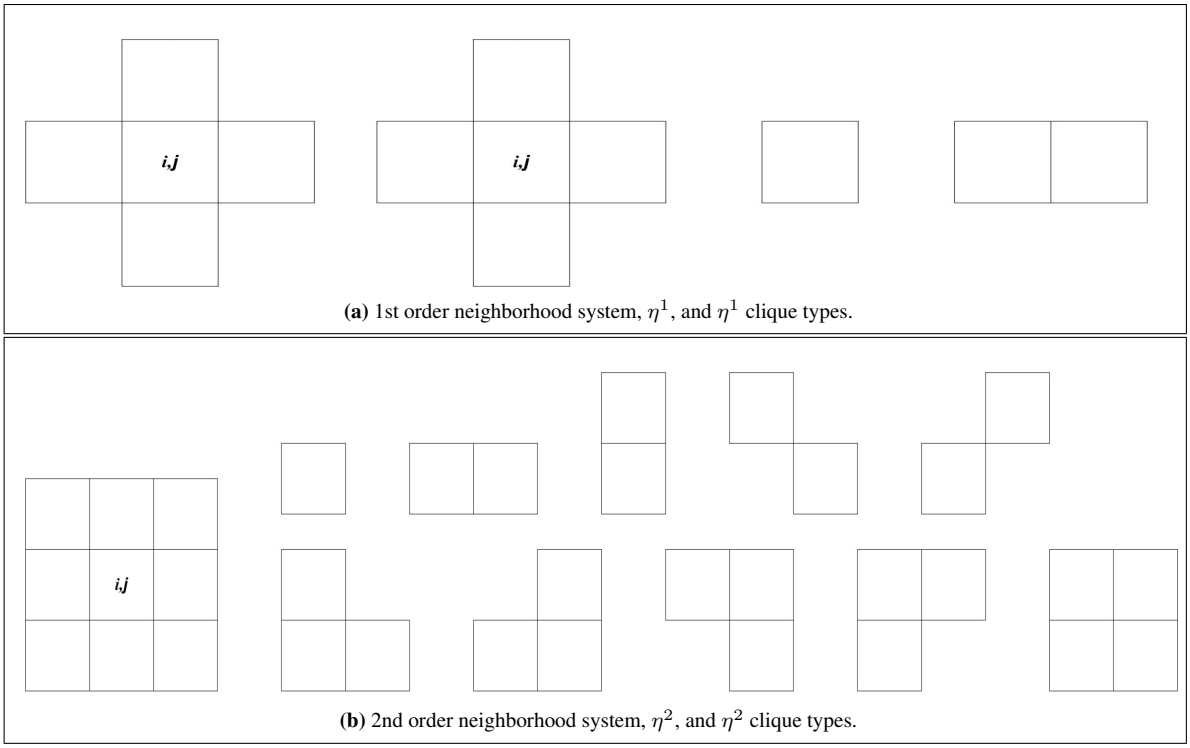
1) $(i,j) \notin \eta_{i,j}$.

2) if $(k,l) \in \eta_{ij}$, then $(i,j) \in \eta_{k,l}$ for any $(i,j) \in L$.

Two types of a neighborhood system are presented in Fig. 2.4. The neighborhood system, n^m, is called the m^{th} *order neighborhood system*. The image pixels at the edges can be ignored or can be modeled with smaller cliques in GD unless the image is assumed periodic.

The associated cliques with a lattice-neighborhood pair (L,η) is defined as follows:

Definition2: A clique of the pair (L,η), denoted by c, is a subset of L such that



(a) 1st order neighborhood system, η^1, and η^1 clique types.

(b) 2nd order neighborhood system, η^2, and η^2 clique types.

Fig. 2.4: Neighborhood systems and their associated clique types [50].

1) c consists of a single pixel, or

2) for $(i,j) \neq (k,l), (i,j) \in c$ and $(k,l) \in c$ implies that $(i,j) \in \eta_{kl.}$.

The collection all cliques of (L, η) is denoted by $C = C(L, \eta)$. *Definition3:* A random field $X = \{X_{ij}\}$ defined on L has Gibbs Distribution(GD) or Gibbs Random Field(GRF) with respect to η if its joint distribution is in the form of

$$P(X = x) = \frac{1}{Z} e^{-U(x)}, \tag{2.40}$$

where

$$\begin{aligned} U(x) &= \sum_{c \in C} V_c(x) \quad \text{energy function} \\ Z &= \sum_{x} e^{-U(x)} \quad \text{associated potential with clique } c, \end{aligned} \tag{2.41}$$

and $V_c(x)$ is the clique potential. $V_c(x)$ which depends on the pixel values in c is the only arbitrary variable, since Z is a normalizing constant. In other words, $V_c(x)$ is the only variable that will be taken into account in optimazion/estimation process.

The physical meaning of the joint distribution in Eq. 2.40 is that the minimum energy function, $U(x)$, can be satisfied by the field which belongs to a same class labels. Although GD is an exponential distribution, a wide variety of distributions for random fields can be formulated as GD by choosing the clique potential function, $V_c(x)$, properly. A more detailed discussion can be found in Besag [46].

2.2.4 Iterated conditional modes

Hence, Besag [51] proposes a deterministic algorithm called "iterated conditional modes" (ICM) which maximizes local conditional probabilities iteratively. It uses the "greedy" strategy in the iterative local maximization, and makes two basic assumptions; one based on the contents of images, and another based on the noise process. The first assumption is that neighboring pixels tend to have the same values because images consist of regions that tend to have roughly the same pixel values except the regions at the edges. There could be seen sharp pixel level changes at the edges in the image. The benefit of this assumption is that it provides an opportunity to change the pixel label, corrupted by noise, by utilizing its local neighborhood information. The second assumption claims that each pixel is corrupted independently, and with some probability, generally considered as Gaussian distribution. In other words, the noise does not corrupt two pixels dependently. If one pixel label is changed by noise, any possible change in its neighbor has again the same probability.

Given the data, I, and the other labels, $y^{(k)}_{S-[i]}$, the algorithm iteratively updates each $y^{(k)}_i$ into $y^{(k+1)}_i$ by maximizing the conditional probability, $P(y^{(k)}_i|I, y^{(k)}_{S-[i]})$, with respect to y_i. From two assumptions stated above, and the *Bayes* theorem, discussed in Section 2.2.1, it follows that,

$$P(y^{(k)}_i|I, y^{(k)}_{S-[i]}) \propto P(y^{(k)}_i|I_i, y^{(k)}_{N_i}) = p(I_i|y^{(k)}_i)P(y^{(k)}_i|y^{(k)}_{N_i}), \tag{2.42}$$

where $y_{N_i}^{(k)}$ denotes the current labeling in the neighborhood, N. Obviously, maximizing a probability of a specific region in the image, $P(y_i^{(k)}|I_i, y_{N_i}^{(k)})$, is more preferable than a probability of the entire image region, $p(y|I)$, since it has less computationally complexity. In addition, maximizing Eq. 2.42 is equivalently minimizing posterior potential, given Eq. 2.43.

$$y_i^{k+1} \leftarrow arg\,\min_{y_i^{(k)}} \quad V(y_i^{(k)}|I_i, y_{N_i}^{(k)}), \tag{2.43}$$

where V is considered as the summation likelihood energy function, found in Eq. 2.34, and potential energy function, shown in Eq. 2.40.

2.3 Hough Transform

The set of all straight lines in an image plane, $x - y$, forms a two-parameter family. An arbitrary straight line can be described by a single point in the parameter space, assuming that the parameter family is fixed. The parametrization, given in Eq. 2.44, defines a straight line by specifying its angle, θ, and its algebraic distance, ρ, from the origin.

$$\rho(\theta) = xcos(\theta) + ysin(\theta), \tag{2.44}$$

where θ is defined in $[0, \pi)$. Every line in $x - y$ plane can be mapped to an unique point in $\theta - \rho$ plane. It is also demonstrated in Fig. 2.5.

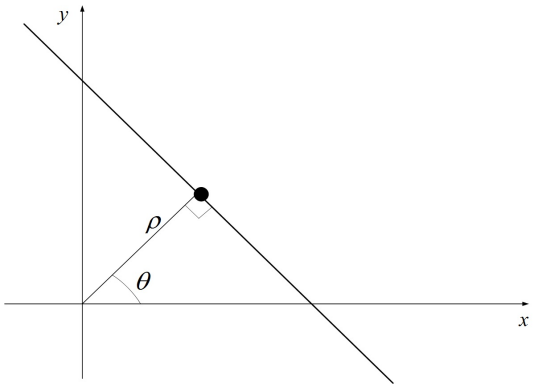


Fig. 2.5: The line parameters in $\theta - \rho$ plane [52].

Suppose that a set of straight lines that fits into a set of $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ of n points wants to be found. The point, (x_j, y_j), is transformed into the sinusoidal curves in the $\theta - \rho$ plane defined in Eq. 2.44 by replacing (x_j, y_j) instead of (x, y). Hence, the problem of detecting collinear points can be converted to the problem of finding concurrent curves [52]. In summary, $1)$ a point in $x - y$ plane corresponds to a sinusoidal curve in the $\theta - \rho$ plane and $2)$ a set of points constitutes a straight line in $x - y$ plane correspond to curves passing through a common point in $\theta - \rho$ plane. Since Hough transform is a reversible process, $1)$ and $2)$ properties run in both ways.

2.4 Run-Length Encoding Algorithm

A run-length encoding (RLE) algorithm has been first introduced by Wahl *et al.* in order to detect long vertical and horizontal white lines in document images [53, 54]. It is improved and utilized to compress the data by Wong *et al.* [55]. The algorithm is supported by bitmap files such as *.tiff*, *.pcx* or *.bmp*. It can be used to extract information from scanned documents to confirm the segmentation of candidate text regions as a post-processing stage [see [56] and [57] for further explanation]. It is also utilized to create databases by assisting in encoding and converting the images/texts in digital documents into computer-processable form. The compression or encoding scheme does not depend on the input's information content. However, its content affects the compression ratio. The significant advantage for RLE algorithm is that it is easy to implement and requires less computation time. Thus, it is preferable to using a complex compression algorithm, or applying no compression technique to an image.

The basic RLE algorithm is applied to a binary sequence in which black pixels are represented by 0's and white pixels by 1's. The binary input sequences, x, is converted into an output sequence, y according to two rules.

1- 0's in the input are changed to 1's in y if the number of adjacent 0's is less than or equal to a threshold C.

2- 1's in x are copied to y exactly.

For instance, with C = 6 the sequence x is mapped into y as follows:

x: 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 1 0 1 0 0 0 1 1,

y: 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1.

To merge into a better segmentation map, it is the best if neighboring black/white zones are linked or separated according to the threshold, C. The threshold value depends on the resolution level. The same technique can also be applied to a document image as a column by column operation, since in some cases the vertical spacing information might be as useful as the horizontal spacing information when determining the text regions.

There are various run-length encoding styles. In a row-by-row operation, the algorithm treats the image as a 1-D data map, rather than as a 2-D data map by starting at the upper left corner and proceeding from left to right across each scan line to the bottom right corner of the map, shown in Fig. 2.6. Alternatively, it can be encoded by starting from the left upper corner and proceeding along the columns, or converting into 2-D tiles, or following a diagonal direction in zig-zag fashion [see Fig. 2.6].

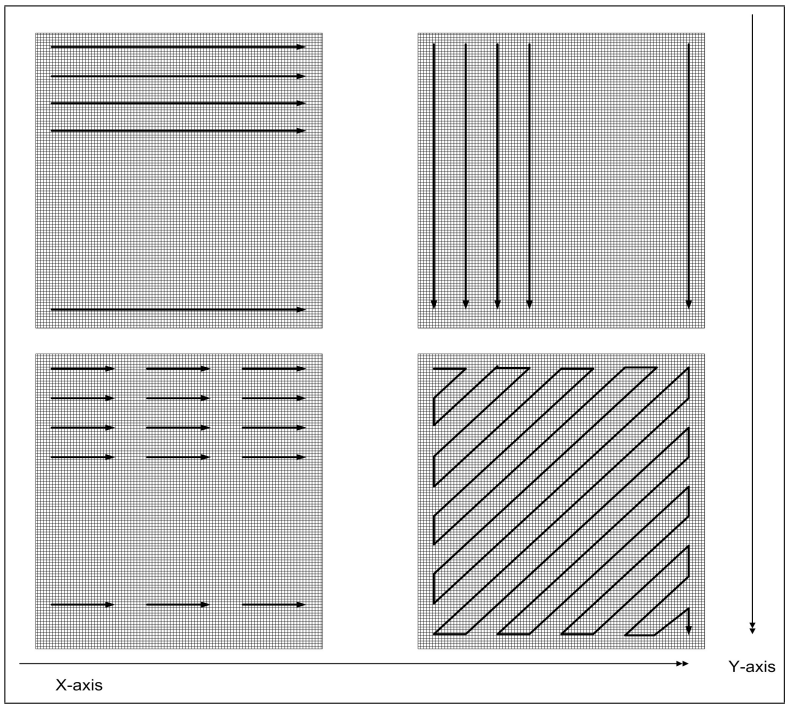


Fig. 2.6: Run length encoding along the X-axis, along the Y-axis, in 2-D tiles and in zig-zag fashion [53].

CHAPTER 3: PROPOSED ALGORITHM

We propose a page layout classification algorithm that takes RGB or grey-scale image as an input. The algorithm starts by a pre-processing module for filtering, image re-sizing, color space transformation, morphological operation and gamma correction which are utilized to limit artifacts because of re-sampling, reduce computation time, eliminate noise, and enhance text characters and illumination effects. Gamma correction is applied to the input grey-level image. If the input image is colored, a color space conversion stage in the pre-processing module transforms the image from RGB color space to CIEL*a*b* space. Then, the lightness channels (L*) is used by the text, photo and line/strong edge detection modules generate three different maps. As a last step, K-Means clustering algorithm is utilized to combine these three maps into one single page layout classification map. A flowchart of the proposed algorithm is shown in Fig. 3.1.

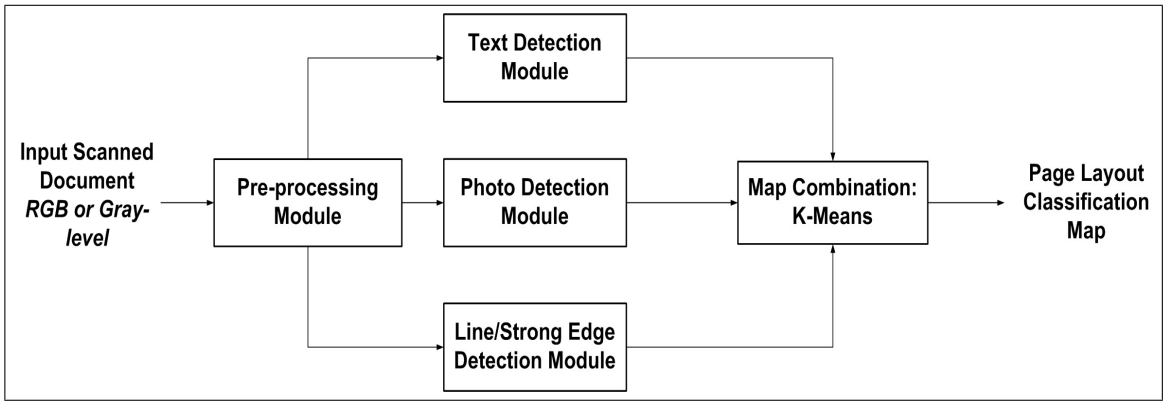


Fig. 3.1: Flowchart of the proposed algorithm.

3.1 Pre-processing Module

This module has different stages for low pass filtering, image re-scaling, color space transformation, morphological operation and gamma correction. The scanned document can be a colored or grey-level image. The objectives of the pre-processing module are to prevent aliasing, reduce computation time, eliminate noise and illumination variations and enhance the edges in text regions. A

block diagram of the module is given in Fig. 3.2.

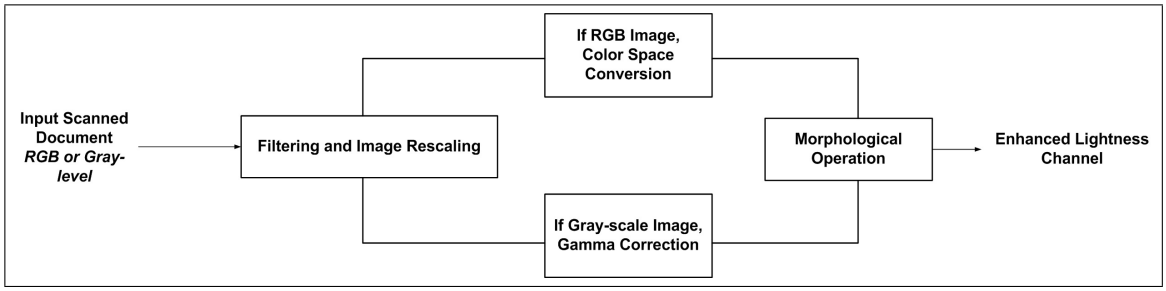


Fig. 3.2: Block diagram of the pre-processing module.

3.1.1 Filtering and image re-scaling

In this study, a typical document size $8,5 \times 11$ inch is used. It is scanned with *300* dots per inch (dpi) which yields an input image of the size 3300×2600 pixels. This technology provides a high resolution image which is advantageous in document classification applications. A drawback however is that it causes the entire process to be computationally expensive. To offset this, the image is down-sampled by a scale factor, $k = 0.25$ using "Bi-cubic interpolation". Before interpolation, 11×11 pixel sized low-pass filter is applied to reduce the effect of ripple patterns that result from aliasing during down-sampling. This limits the impact of aliasing on the output image, and minimizes the artifacts that might occur. In bi-cubic interpolation, the output pixel value is computed by weighting the average of the pixels in the nearest *4-by-4* neighborhood. The scheme of image re-scaling step is presented in Fig. 3.3.

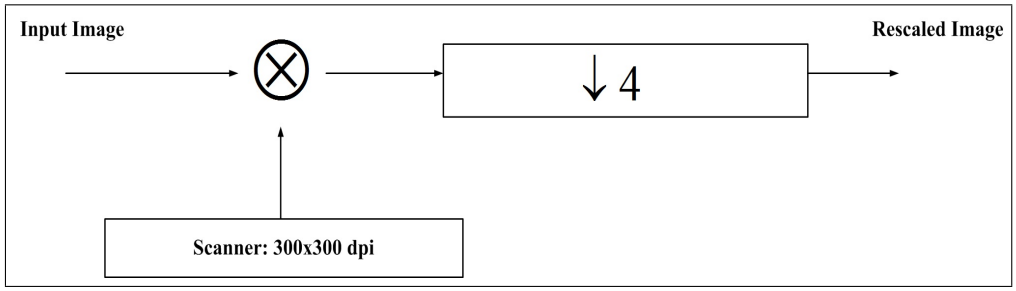


Fig. 3.3: Scheme of image re-scaling.

3.1.2 Color space transformation

In 1976, the Commission International de lÉclairage (CIE) proposes CIEL*a*b* color space which is a uniform color space, to model the human perception of color, and to provide a standard scale for comparison of color values. This color space system is often used in the quality control of colored products since it is based on human color perception [58]. For instance, if the color of a production sample is detected in the CIEL*a*b* color space, color differences in the production sample can be compared with the predetermined standards.

In a uniform color, the differences between points plotted in the color space correspond to visual differences between colors plotted. It is designed in a cube form. The L* axis lies from top to bottom between 0, representing black, and 100, representing a perfect light diffuser. The a^* and b^* do not necessarily lie between specific numbers. $+a$ and $-a$ is red and green. $+b$ and $-b$ is blue and yellow. The colors in the color space can be considered as the combinations of red and yellow, red and blue, green and yellow, and green and blue. 3-D coordinate system is introduced in order to determine the exact combination of these colors in a product.The configuration of the coordinate system is depicted in Fig. 3.4 below.

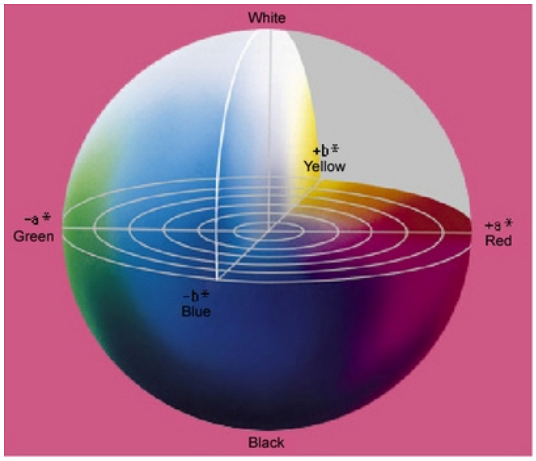


Fig. 3.4: CIEL*a*b* color space [58].

Color space transformation is applied to the original scanned document when it is scanned in RGB color space, as presented in Fig. 3.2. One of the benefits of this color space transformation is

to provide perceptual uniformity. It conforms to many digital image manipulations better than the RGB space in image sharpening and artifacts removal. Besides this, the color components (a* and b*) can be isolated by employing only L* component since CIEL*a*b* color space transformation separates the color and lightness information.

In this study, the down-sampled image is transformed to the CIEL*a*b* color space where the only L* component is used in the proposed algorithm. The color transformation which is described from Eq. 3.1 to Eq. 3.3, can be achieved by first transforming the image from RGB to CIEXYZ space, given in Eq. 3.1, and then from CIEXYZ to CIEL*a*b* space as shown in Eq. 3.2 and 3.3. This transform is based on ITU-R Recommendation BT.709 using the D-65 white point reference. The error in transforming from RGB to CIEL*a*b is approximately 10^{-5}.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412 & 0.357 & 0.180 \\ 0.212 & 0.715 & 0.072 \\ 0.019 & 0.119 & 0.950 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{3.1}$$

After RGB to CIEXYZ conversion is performed, the components of the CIEL*a*b* color space can be computed by

$$\begin{aligned} L &= 116 f(Y/Y_n) - 16 \\ a &= 500[f(X/X_n) - f(Y/Y_n)], \\ b &= 200[f(Y/Y_n) - f(Z/Z_n)] \end{aligned} \tag{3.2}$$

where

$$f(t) = \begin{cases} t^{1/3} & \text{if} \quad t > \left(\frac{6}{29}\right)^3 \\ \frac{1}{3}\left(\frac{29}{6}\right)^2 t + \frac{4}{29} & otherwise \end{cases}. \tag{3.3}$$

In Eq. 3.3, X_n, Y_n and Z_n are the CIEXYZ color space tristimulus values of the D-65 white point reference.

3.1.3 Gamma correction

The RGB to CIEL*a*b* color space conversion has an inherent gamma correction. Therefore, to simulate similar behavior for gray-scale scanned document, a gamma correction process is applied. It is performed on gray-level scanned documents to eliminate their illumination variances, and to suppress the noise at the background region. In general, the gamma correction process, a nonlinear operation as shown in Eq. 3.4, takes linear light information (video or still imagery) and changes it into a display more harmonious with the way the eye actually processes information. The aim of the gamma correction is to create a realistic image in terms of shading, intensity, luminance and/or brightness. Plots of the Eq. 3.4 for various values of γ are demonstrated in Fig. 3.5.

$$I_{out} = c(I_{in})^{\gamma}, \tag{3.4}$$

where I_{out} and (I_{in}) are the output image and input image and c is a constant, generally considered as 1. The gamma factor, γ, is taken as 2.2 in this study.

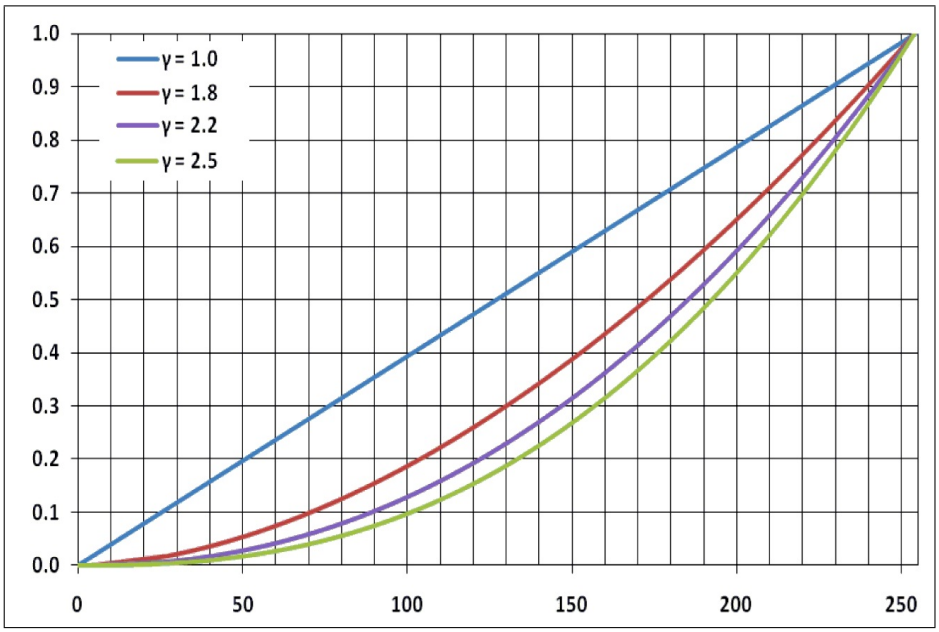


Fig. 3.5: Plots for various values of γ.

3.1.4 Morphological operations - dilation

A pre-processing module based on morphological dilation operation is employed to enhance high-frequency (edge) regions in the enhanced L* component of the map. The dilation operation scans the input intensity image to find local maxima in a given direction over a small window. The operation is employed twice to emphasis high-frequency regions in horizontal and vertical directions. These two maps ($Dilation_{Horizontal}$ and $Dilation_{Vertical}$) are averaged and subtracted from the input L* channel as shown in Eq. 3.5. The $|.|$ sign stands for absolute value in the equation.

$$Enhanced\, L^* = \left| L * -\frac{1}{2}(Dilation_{Horizontal} + Dilation_{Vertical}) \right|. \tag{3.5}$$

3.2 Text Detection Module

The text detection module uses the enhanced lightness channel (L*) as described in the pre-processing module. Multiple-scale wavelets decomposition is applied and local energy (a variable size window operation) is computed in high frequency sub-images as horizontally, vertically, and diagonally. The energy maps are up-sampled to the size of the input enhanced lightness channel, and averaged to generate a text-candidate map. Finally, a module based on RLE is utilized to verify text regions in the text-candidate map. These sub-modules are detailed in the following sections.

3.2.1 Wavelet decomposition and energy sub-module

The goal of this operation is to identify the candidate text regions. We employ a basic assumption which is that text regions have high variation in a small neighborhood area, in addition to contrasting with a background. This addresses the most extreme case of having text with a small font size and a background with too little contrast. The proposed algorithm also handles more complicated scanned documents where have complex background as exhibited in the Chapter 4.

The proposed technique starts by applying wavelets decomposition to the $Enhanced\, L^*$ generated by using Eq. 3.5. The DWT methodology is utilized where multiple levels are applied to the

low-frequency approximated sub-image as shown in the block diagram in Fig. 3.6. The energy is computed using a variable-size sliding window. The window size varies in relation to the original document spatial size, and to the wavelets decomposition level. Notice that the average value of the wavelets coefficients in the given window has been subtracted from all coefficients. This is to eliminate any bias in energy values that could be caused by lighter grey-scale background. In other words, it is performed to emphasize the contrast between the text and background regions to prevent the color of the background region from dominating the energy values. This procedure is applied twice where first the local average of the neighboring coefficients (I'_{local}) is used and then the global average of all coefficients (I'_{global}) in the sub-image is used in the other term of Eq. 3.6.

$$Text_{Energy\,Map} = \sqrt{\sum_{x,y \subset W} \left(I(x,y) - I'_{local}\right)^2} + \sqrt{\sum_{x,y \subset W} \left(I(x,y) - I'_{global}\right)^2}, \tag{3.6}$$

where W stands for the local window and $I(x,y)$ is the wavelets coefficient at the location x and y.

Fig. 3.6 shows a block diagram for the wavelets decomposition and energy maps sub-module. Two DWT levels using *Daubechies* 4-tap filter-banks are shown in the figure. However, up to four levels of DWT can be applied depending on the spatial size of the original scanned document. The range of the energy maps is normalized before up-sampling them to the original document size and generating their average map. Bi-cubic interpolation [see Section 3.1.1] technique is used to resize the energy with the scaling factor (2^s) where s is the wavelet scale (level).

The purpose of the wavelets and energy computation is to generate a text-candidate map that outlines the exact text-candidate regions. However, these operations generate grey-scale maps that signify high-frequency regions such as text, texture, and edges. Therefore, a thresholding operation based on Otsu's method [59] is employed to generate the binary map. Some of the target scanned document in our test database have text written using deferent colors or gray-levels where they yield different energy levels. Hence, applying histogram (intensity value) adjustment operation before the thresholding stage helps to reduce the energy variation due to text color or gray-level

difference. Another image enhancement operation is applied to the binary text-candidate regions as a post-processing process by removing any region with insignificant size less than 0.03% of scanned document size.

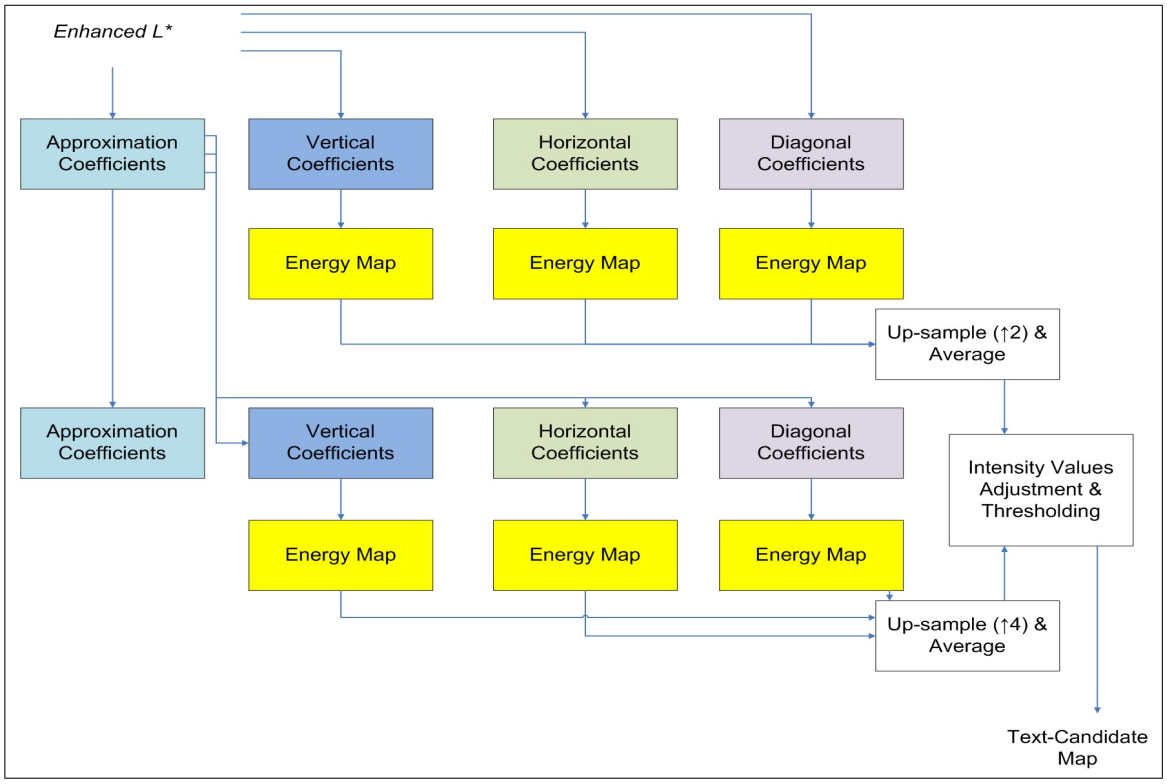
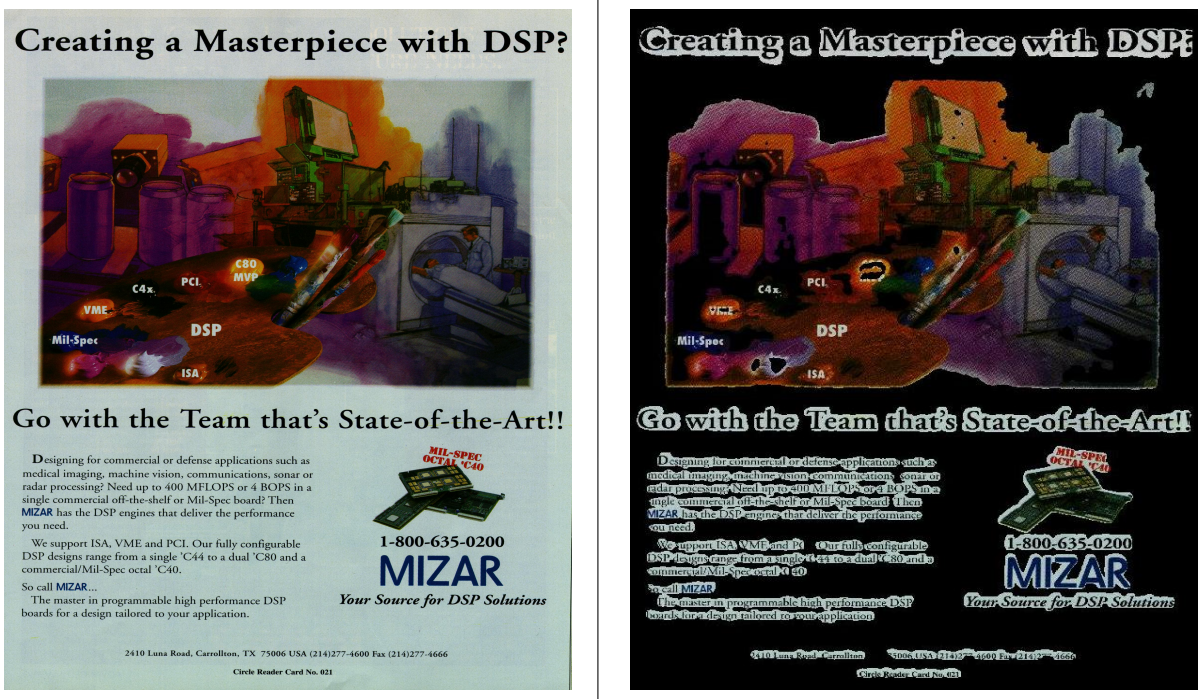


Fig. 3.6: Block diagram of the wavelet decomposition and energy maps sub-module.

A illustration for this step is demonstrated in Fig. 3.7. Almost entire text region, corresponded to true positives, is successfully detected in this step as observed in Fig. 3.7(b). Additionally, background region manages to be extracted from the map, as well. Nevertheless, big portion of the photo region cannot be excluded from the text map since high frequency content is utilized in this step. To eliminate the photo region from the text map, a validation stage called Text region confirmation is utilized to obtain more accurate text map.



(a) Original image.

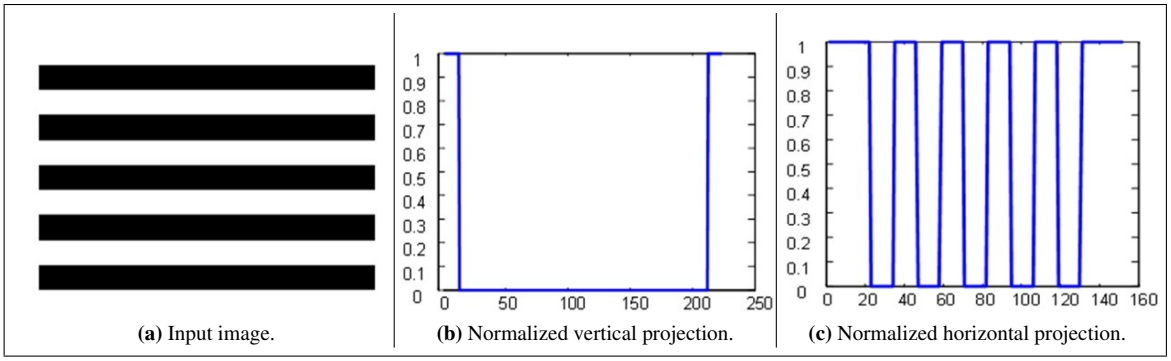
(b) Initial text map after wavelet analysis.

Fig. 3.7: Initial text region classification.

3.2.2 Text region confirmation

This module uses the text-candidate maps generated in Section 3.2.1 and the enhanced L* channel of the original scanned document. It is assumed that a text region is typed in a line or multiple lines or paragraphs format. If any text-candidate region is considered by itself, its structure should follow this assumption. That is, it generates a set of peaks and valleys if averaged in horizontal or vertical direction (at least in one direction). The characteristics of these peaks and valleys indicate the font size used in the written text and the distances between the lines.

Fig. 3.8 shows an example of vertical and horizontal projections of a text region. These projections are normalized by the image height and width, respectively. The RLE algorithm is applied to the projection vectors and the mean and standard deviation (SD) are computed. If the paragraph is written in a consistent font and the spacing between the lines is fixed, this will generate a relatively low SD value in comparison with the average line width which indicates a text region. Therefore,

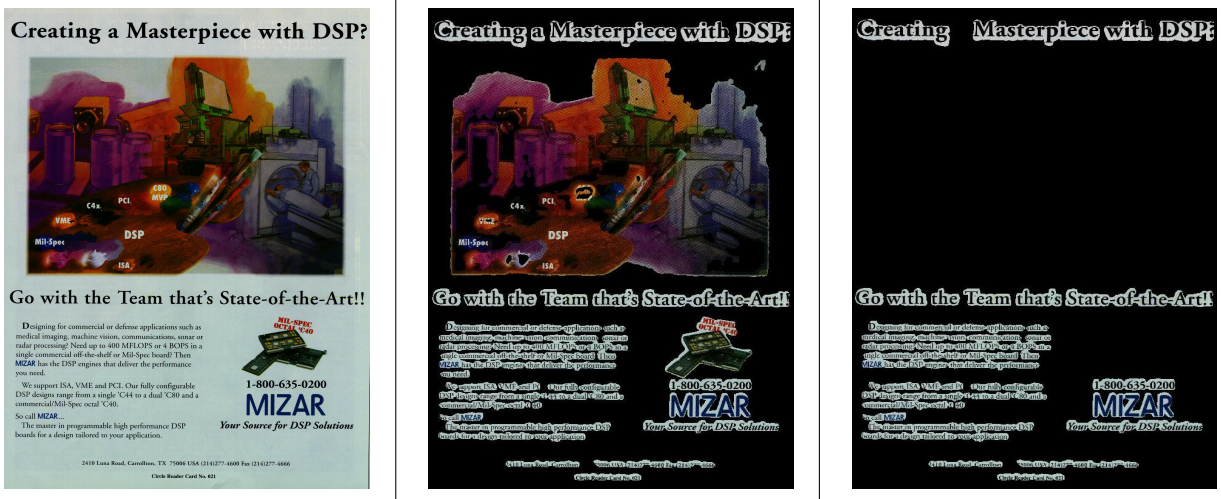


(a) Input image. **(b)** Normalized vertical projection. **(c)** Normalized horizontal projection.

Fig. 3.8: Example of vertical and horizontal projections of text region.

if the average line width is higher than the variation (SD) at least in one direction, the image region is identified as a text region. This is given that there is a pattern (peaks and valleys) at least in one of the projections in Fig. 3.8, otherwise, it is not a text region.

In Fig. 3.9, final text classification map is presented with one of the scanned documents which is used to evaluate the performance of the algorithm. Obviously, the main body of the photo region is well-segmented and omitted from the final text map [see Fig. 3.9(c)] as compared to the intermediate map in Fig. 3.9(b). Additionally, although there are many separate text zones in different font-sizes, they are classified accurately which shows that text detection module is font-size independent.



(a) Original image. **(b)** Initial text map after wavelet analysis. **(c)** Final text map after RLE.

Fig. 3.9: Intermediate results of the algorithm.

3.3 Photo Detection Module

Similar to the text detection module, the enhanced lightness channel (L*) is used as an input for the photo detection module. The broken characters in the text regions are enhanced by morphological operation to bridge the gaps between the text characters. The enhanced image is initially segmented into three classes, background, text and photo, by employing basis vectors with projection method. After initial segmentation is achieved, MRF-MAP with ICM is applied to utilize contextual information and merge a more accurate photo map. As a last step, missing blocks (false positives) which are fully surrounded by detected image block(s) are included in the final photo map. A block diagram of the photo detection module, where its sub-modules are explained in the following sections, is drawn in Fig. 3.10.

Fig. 3.10: Block diagram of photo detection module.

3.3.1 Block-wise segmentation based on basis vectors projection

Block-wise segmentation based on projection basis vectors is first proposed in [39]. However, the raw image is used as an input in his study. In other words, no pre-processing technique is applied to the input RGB color image. Additionally, the technique introduced in [39] determines the optimum block-size by utilizing alternating blanks between words and text lines. The constraint for an optimum block size is to include sufficient information number of text lines in the block. On the contrary, the block size is fixed to 32×32 pixels in this study. If there is any sizing operation applied to the input image in the pre-processing module, an updated block pixel size, denoted by $B \times B$, is obtained by multiplying the 32×32 window with the scaling factor. Besides this, the decision criterions for background, text and photo are modified to yield a more robust segmentation map for different types of scanned documents which have complex color background.

Initially, block-wise segmentation is achieved by utilizing projection basis vectors. These different types of basis vectors represent either text, background or image regions. The image is divided into $S \times S$ non-overlapping blocks. For each block, the gray levels are horizontally projected in order to constitute a row-vector, $P = [p(0), p(1) \cdots, p(31)]^T$ where it represents the projection values for horizontal line in selected block. $p[n]$, given in Eq. 3.7 takes either $+1$ or -1 depending on whether n^{th} line is the text or background. $+1$ and -1 are for text and background respectively.

$$p[n] = \begin{cases} +1 & \text{if } \left[\sum_{j=0}^{S-1} I\{\Psi(i,j)\}\right] > 32 \times T_2 \\ -1 & \text{if } otherwise \end{cases}, \tag{3.7}$$

where

$$I\{\Psi(i,j)\} = \begin{cases} +1 & \text{if } \Psi(i,j) > T_1 \\ 0 & \text{if } otherwise \end{cases}, \tag{3.8}$$

where Ψ is the image and $I\{\Psi\}$ is the corresponding binarized image. First, the image is binarized with thresholding method according to the threshold value, T_1. Then, if the corresponding line, n, represents a text line, then most of the pixels at that line in the block takes lower values than T_1. Otherwise, it takes higher values and indicates background. Eq. 3.8 can be interpreted as binarization formula for Eq. 3.7. Then the number of 1's are computed in each horizontal line of the block, $p[n]$, to determine whether the corresponding line is a text or background [see Eq. 3.7].

The basis vectors, shown in Fig. 3.11, are introduced to decide which class the corresponding block belongs to. It is apparent that any of the two basis vectors in Fig. 3.11 are not only orthogonal to each other but also orthonormal as formulated in Eq. 3.9,

$$\langle \Phi_i, \Phi_j \rangle = \sum_{k=1}^{8} \phi_{ik}\phi_{jk} = 0 \qquad \langle \Phi_i, \Phi_i \rangle = \sum_{k=1}^{8} \phi_{ik}\phi_{ik} = 1, \tag{3.9}$$

where $1 \leq i, j \leq 8$ for $\forall i, j$ and $\langle . \rangle$ represents the inner product of any two basis vectors. Φ_1 and Φ_2 represents a background/image region patterns and the rest represents a text region patterns.

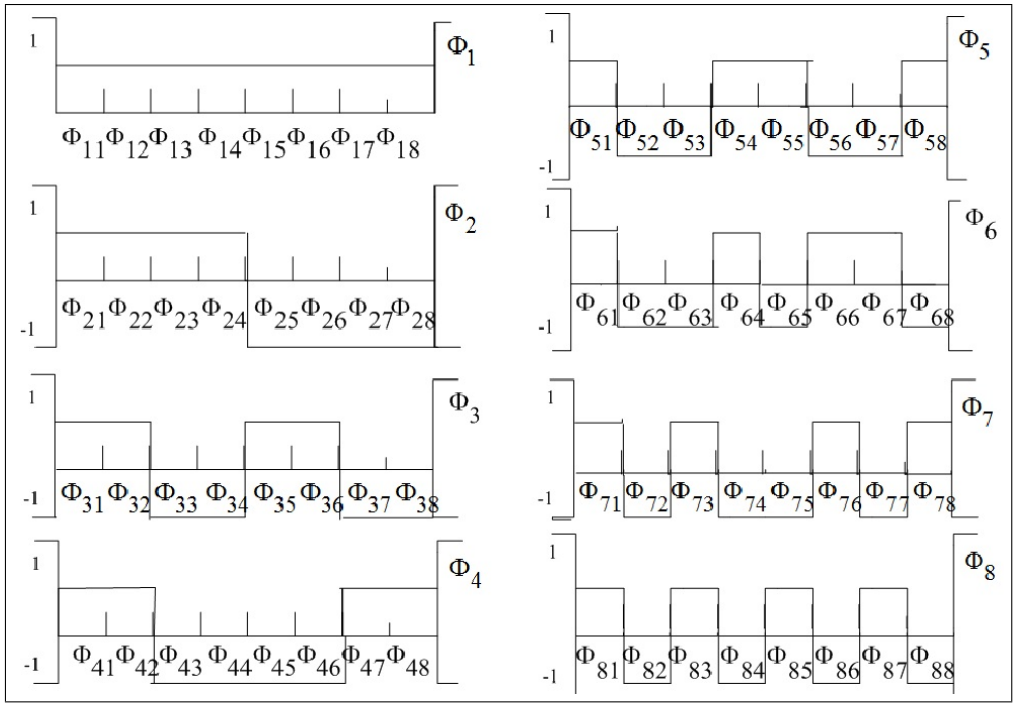


Fig. 3.11: Basis vectors for the determination of the best fit for the region in the block [39].

$$B = [\Phi_1, \Phi_2 \cdots \Phi_8]\,. \tag{3.10}$$

Eight basis vectors are generated and projection vector, P, is re-arranged to multiply with 8×8 basis matrix, B, given in Eq. 3.10. In other words, P vector is grouped by k to yield 1×8 vector by using Eq. 3.11 given below.

$$P^{'} = \sum_{n=1}^{l+1\times S/8} p[n], \tag{3.11}$$

where $l = 1, \cdots, 8$. $P^{'}$ becomes equal to $P^{'} = [\,p^{'}[1], \cdots, p^{'}[8]\,]^T$ for initial block segmentation purpose. Then, the re-organized vector, $P^{'}$ is represented by using basis vectors with some weighting coefficients. $P^{'}$ can be re-written by using the basis matrix and weighting coefficients, presented in Eq. 3.12.

$$P^{'} = [a[1]\Phi_1 + a[2]\Phi_2 + a[3]\Phi_3 + a[4]\Phi_4 + a[5]\Phi_5 + a[6]\Phi_6 + a[7]\Phi_7 + a[8]\Phi_8]\,, \tag{3.12}$$

where $a[n]$ represents the weighting coefficient of how good $P^{'}$ fits into the corresponding basis vector, Φ_n. Moreover, $a[n]$ is obtained by the inner product of $P^{'}$ and Φ_n as shown in Eq. 3.13.

$$a[n] = \left\langle P^{'}, \Phi_n \right\rangle = p^{'}[1]\Phi_1 + p^{'}[2]\Phi_2 + \cdots + p^{'}[8]\Phi_8. \tag{3.13}$$

To finalize the initial block segmentation, class labels, x, are assigned as 2 for background, 1 for image, and 0 for text according to Eq. 3.14.

$$x = \begin{cases} 2 & \text{if } |a[1]| \geq |a[m]| \quad \text{and} \quad |a[1]| > \sum_{i=2}^{8} |a[i]| \quad \text{and} \quad a[1] > 0 \\ 1 & \text{if } |a[1]| \geq |a[m]| \quad \text{and} \quad |a[1]| > \sum_{i=2}^{8} |a[i]| \quad \text{and} \quad a[1] < 0\,, \\ 0 & \text{if} \qquad\qquad\qquad\quad otherwise \end{cases} \tag{3.14}$$

Where $m = 2, \cdots, 8$. If $|a[1]| \geq |a[m]| \quad \text{and} \quad |a[1]| > \sum_{i=2}^{8} |a[i]|$, then the first coefficient, $a[1]$, becomes the most dominant coefficient. In addition to this condition, if $a[1] > 0$, the pixels values in the selected block tend to have monotone levels, corresponding to a background region. On the other hand, if $|a[1]| \geq |a[m]| \quad \text{and} \quad |a[1]| > \sum_{i=2}^{8} |a[i]|$, then the first coefficient again becomes the most dominant coefficient. However, in this case, if $a[1] < 0$, then the pixels values in the selected block tend to have both monotone and non-white levels which represent a image region. Otherwise, the block belongs to a text region which consists of a set of horizontal gaps between the words.

A demonstration of this stage is presented in Fig. 3.12. Notice that, although there are some false detections, main body of the photo zone (entire image) are well-extracted which ensures fairly accurate detection rate for the next phases of the module. Additionally, the stage overcomes

the reflection at the background and eliminates from the actual document after utilizing the pre-processing module explained explicitly in Section 3.1. However, there are some false positives that fail to be included in photo map at this stage. For instance, lighter pixels on balloons [see Fig. 3.12(a)] are detected as background presented in Fig. 3.12(b). Besides this, text regions are also eliminated appeared under the printer images in Fig. 3.12(b) since their patterns matches with the projection vectors which represents a text region pattern.

(a) Original image.

(b) Initial map after block-wise segmentation.

Fig. 3.12: Block-wise segmentation based on basis vectors projection.

3.3.2 Markov random field: MAP segmentation

In basis vectors projection based segmentation, contextual information is not considered. This yields some false detection. To eliminate these blocks, block-based MRF-MAP image segmentation algorithm is employed. Additionally, the class label field, $X = x$, is assumed to be MRF model and ICM is utilized to increase the convergence rate. The aim is to find the expression in Eq. 3.15.

$$\arg\max_{y} p(y|x) \propto \arg\max_{y} \frac{p(x|y)p(y)}{p(x)}, \tag{3.15}$$

where $p(y)$ is a-priori probability assumed to be Gibbs distribution, shown in Eq. 3.16, $p(x)$ is the $S \times S$ block, selected from the image which can be ignored since it is deterministic and $p(x|y)$ is

conditional probability distribution function (pdf).

$$p(y) = \frac{1}{Q} exp\left\{ -\sum_{c \in C} V_c(y) \right\}, \tag{3.16}$$

where C is the set of cliques in $S \times S$ block, Q is the Gibbs constant and $V_c(y)$ is shown in Eq. 3.17.

$$V_c(y) = \begin{cases} \beta & \text{if the class labels in the pair clique are different} \\ -\beta & \text{otherwise} \end{cases}. \tag{3.17}$$

In Eq. 3.17, β is a constant for clique potential which is chosen to be 1.6. Each $S \times S$ non-overlapping block is assumed to be independent and have a Gaussian distribution. To find the conditional pdf for the image, a formulation is required which is given below in Eq. 3.18 [see Section 2.2.2 for detailed explanation].

$$\begin{aligned} p(Y = y|X = x) &= \prod_{i \in I} p(Y_i = y_i|X_i = x_i) \\ &= \prod_{i \in I} \frac{1}{\sqrt{2\pi\sigma_{x_i}^2}} exp\left\{ \frac{-(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} \right\}, \end{aligned} \tag{3.18}$$

where $\mu_{x_i}, \sigma_{x_i}^2$ and I are the mean, variance and the given image, respectively. They are sufficient statistical features to formulate the blocks. After defining each variable, by using Eq. 3.16 and 3.18, the expression in Eq. 3.15 can be re-written as follows;

$$\begin{aligned} E_{max} &= max\left\{ p(Y = y|X = x)p(Y = y) \right\} \\ &= max\left\{ \left(\frac{1}{\sqrt{2\pi\sigma_x^2}} \right)^{S \times S} exp\left\{ -\sum_{i=1}^{S} \sum_{j=1}^{S} \frac{[y(i,j) - \mu_x^2]^2}{2\sigma_x^2} \right\} \times \frac{1}{Q} exp\left[-\sum_{c \in C} V_c(y) \right] \right\}. \end{aligned} \tag{3.19}$$

In Eq. 3.19, $\left(\frac{1}{\sqrt{2\pi\sigma_x^2}} \right)^{S \times S}$ and $\frac{1}{Q}$ can be omitted from the expression since they are constant and have no effect in optimization process. In addition, if $(-)$ sign is combined with maximization argument, it becomes equal to minimization operation as given in Eq. 3.20.

$$E_{max} = max\left\{exp\left[-\sum_{i=1}^{S}\sum_{j=1}^{S}\frac{[y(i,j)-\mu_x^2]^2}{2\sigma_x^2}\right]\times exp\left[-\sum_{c\in C}V_c(y)\right]\right\}$$
$$= min\left\{exp\left[\sum_{i=1}^{S}\sum_{j=1}^{S}\frac{[y(i,j)-\mu_x^2]^2}{2\sigma_x^2}\right]\times exp\left[\sum_{c\in C}V_c(y)\right]\right\}. \quad (3.20)$$

Then, take the $ln()$ of both sides, this yields to Eq. 3.21.

$$E = min\left\{\sum_{i=1}^{S}\sum_{j=1}^{S}\frac{[y(i,j)-\mu_x^2]^2}{2\sigma_x^2}+\sum_{c\in C}V_c(y)\right\}. \quad (3.21)$$

where the first term corresponds to the constraint region intensity to match available data and the second one imposes spatial continuity. The formula in Eq. 3.21 computes the energy over the entire image which is computationally time consuming. Hence, ICM is applied to minimize the computation time while the algorithm performance is maintained. Instead of considering the entire image which is represented by $\sum_{i=1}^{S}\sum_{j=1}^{S} exp[\cdots]$, the energy term, E, is computed for the neighborhood pixels/blocks. Hence, Eq. 3.21 takes the form in Eq. 3.22. The detailed discussion about ICM can be found in Section 2.2.4.

$$E = min\left\{\sum_{i\in\zeta^m}\sum_{j\in\zeta^m}\frac{[y(i,j)-\mu_x^2]^2}{2\sigma_x^2}+\sum_{c\in C}\sum_{y\in\zeta^m}V_c(y)\right\}, \quad (3.22)$$

where the pixel, (i,j) is the center pixel of the given neighborhood system, ζ^m. It represents the pixels in m^{th} order neighborhood system. "$2nd$ order neighborhood clique system" is used in ICM iterations. This iterated approach is executed until the convergence condition, denoted by CC and given below, is satisfied. Current class labels are updated with the following steps;

1) For given current class labels, x, calculate (μ_0,σ_0^2), (μ_1,σ_1^2) and (μ_2,σ_2^2) which are the mean and variance of text, image and background zones.

2) Compute E, given in Eq. 3.19, for each block in the image and update current class labels of the blocks, x', by selecting the class $(0,1$ or $2)$ which maximizes E.

3) If the $CC = \frac{\sum_{i=1}^{B}\sum_{j=1}^{B} sign[x'(i,j)-x(i,j)]}{B^2} < T = 0.1$, stop. Otherwise, go to step 1.

where

$$sign\left[x'(i,j) - x(i,j)\right] = \begin{cases} 1 & \text{if } x'(i,j) - x(i,j) \text{ are different} \\ 0 & \text{otherwise} \end{cases}. \tag{3.23}$$

Note that, only if $x'(i,j)$ and $x(i,j)$ have different class labels, the result of the summation will change, otherwise no effect will be seen. The algorithm usually converges in $2-3$ iterations.

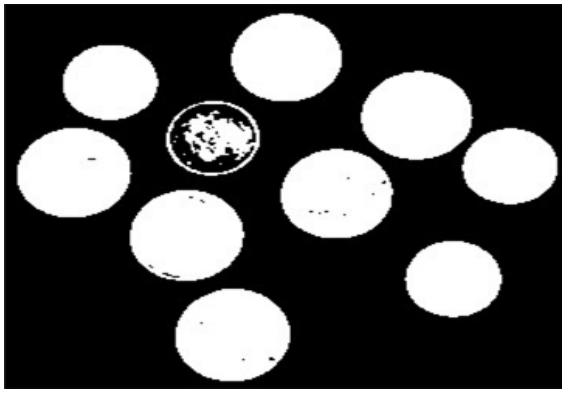
After MRF-MAP segmentation, the original image, and intermediate photo maps are illustrated in Fig. 3.13. Many false detections (false positives) in the scanned document especially the text regions in many locations are included in photo classification map by utilizing the contextual and spatial location information with MRF-MAP optimization. It is worth noticing that, text regions under the printer images and at the bottom of the document are detected successfully although they are mis-classified at the previous step [see Fig. 3.13(b) and (c)]. However, some regions cannot be still included in the photo map since they are large enough to force the MRF-MAP optimization technique an error.

(a) Original image **(b)** Initial map **(c)** Intermediate map after MRF-MAP

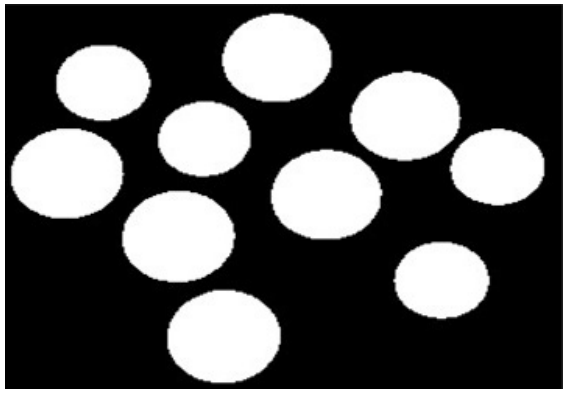
Fig. 3.13: Segmentation maps before post-processing.

3.3.3 Photo map enhancement process

After MRF-MAP optimization segmentation is completed, photo map enhancement step is performed to eliminate false negatives (black pixels in Fig. 3.14(a)) which are surrounded by classified blocks (white pixels in Fig. 3.14(a)) and to merge final photo map as it is illustrated in Fig. 3.14(b). For connectivity, second-order neighboring system is used. This stage consists of several dilation operations and it continues to iterate until the contour of the initial sub-image (the rings with black pixels in Fig. 3.14(a)) fits under a main detected sub-image (the rings in Fig. 3.14(b)). The process stops when further dilation causes changes at the shape (contour) of the main detected image.



(a) Before photo map enhancement process.

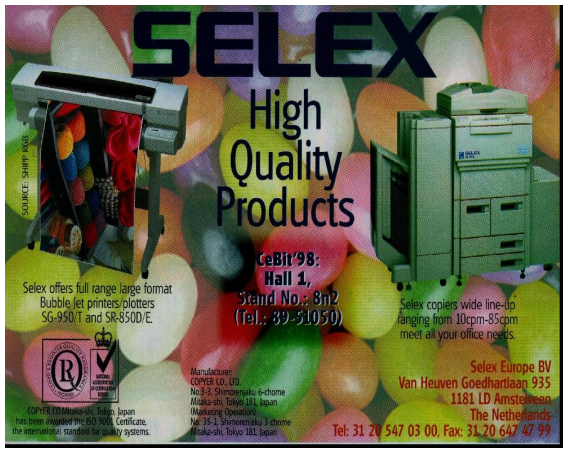


(b) After photo map enhancement process.

Fig. 3.14: Photo map enhancement process.

Fig. 3.15 exhibits the effect of the stage on photo map. The false negatives that cannot be detected in the previous stages are well-classified and included in the photo map. Although MRF-MAP segmentation technique misses these regions since the neighboring blocks do not provide enough information to segment them as a photo, the enhancement stage achieves to detect these false negatives and yield the module an accurate photo map [see Fig. 3.15(a) and (b)].

(a) Intermediate photo map after MRF-MAP



(b) Final photo map after enhancement

Fig. 3.15: Segmentation maps for MRF-MAP and enhancement process.

3.4 Strong Edge / Line Detection Module

Lines are detected in the proposed algorithm using Hough transform. It starts by employing Canny edge detection methodology to generate an edge map of the input enhanced L* channel. Next, the Standard Hough transform (SHT) is applied where the parametric representation of a line, given in Eq. 2.44, is used.

Applying the Hough transform for all edge points in the edge map generates a parameter space matrix whose rows and columns correspond to ρ and θ, respectively. Peak values in this space represent potential lines in the input image. Several parameters that are essential for the success of the line detection algorithm are set empirically based on the test data-set. They are as follows:

1) A threshold value equals to 20% of the maximum peak is used to identify potential lines.

2) The maximum number of peaks to identify in parameter space matrix is set to *30*.

3) A scalar value that specifies whether merged lines should be kept or discarded. Lines shorter than *300* pixels are discarded.

4) A scalar value that specifies the distance between two line segments associated with the same Hough transform bin. When the distance between the line segments is less than *15* pixels, the

Hough methodology merges the line segments into one single line segment.

It is critical to identify strong edges such as outlines of objects in scanned documents to ensure a seamless transition when applying different image enhancing techniques to neighboring regions. Edges are ideal locations to embed the transition boundary, for example different color quantization tables for memory color regions such as sky and grass. Imagine an image in which sky and grass meet at the horizon. Enhancing these memory colors (blue for sky and green for grass) separately, and implanting the transition region over the strong edge (horizon line) in the image would enhance the image's overall visual quality, while minimizing any fault caused by the color correction process.

The strong-edge detection technique uses the edge map, generated by the Canny edge detection algorithm, as an input. Edge pixels are linked together into lists of sequential edge points, one list for each edge contour. A contour or edge-list starts/stops at an ending or a junction with another contour/edge-list. A thresholding technique is utilized to eliminate short edges where contours less than *200* pixels long are discarded.

3.5 Map Combination

In this module, train maps for text and image regions are determined in order to obtain features by utilizing text and photo maps which are obtained in the text detection and photo detection module. Besides train maps, an intersection map, common regions in both text and photo map, is calculated as shown in Eq. 3.24. Features are then extracted by utilizing train maps to characterize the images in the intersection map. They are selected to maximize the margins between two classes. A block diagram of the map combination module is shown in Fig. 3.16. A demonstration of the module is also exhibited in Fig. 3.17 where the rectangular, shown in cyan color in Fig. 3.17(a), is the intersection region which is merged into photo, text or both maps.

The regions in blue and green in Fig. 3.17 are the train maps for photo and text regions which are used for feature extraction. After the features are computed, K-Means clustering algorithm is

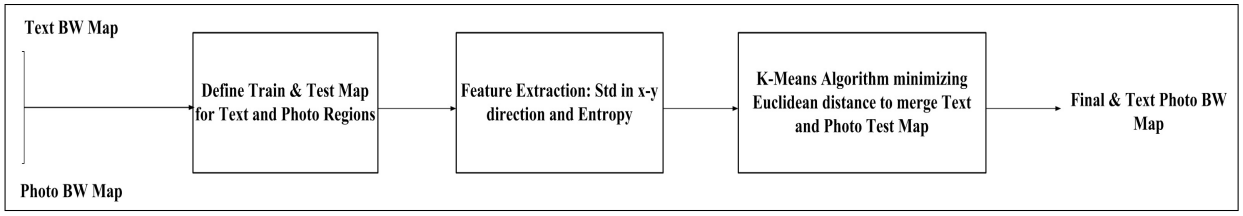
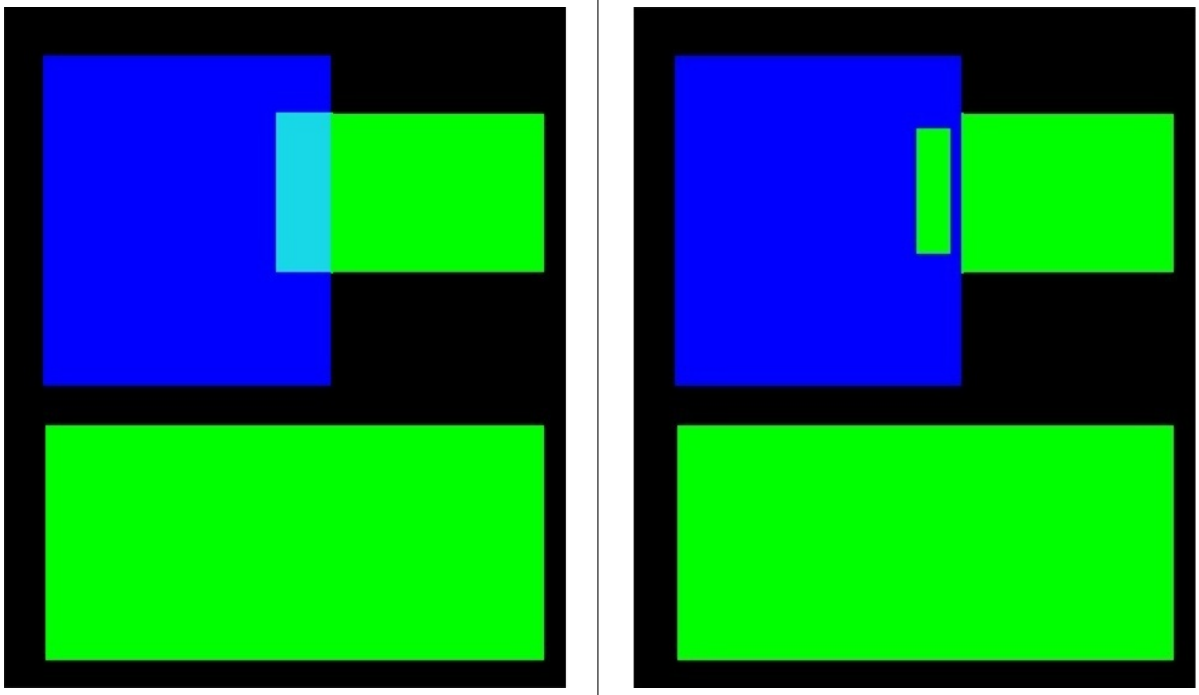


Fig. 3.16: Block diagram of the map combination module.

applied by minimizing the Euclidean distance to merge the final text and photo map, illustrated in Fig. 3.17(b). This module is skipped if there is no intersection map, or the data in the train maps is not sufficient to compute the features.



(a) Before map combination process

(b) After map combination process

Fig. 3.17: Map combination process

3.5.1 Training/Testing maps for text and photo regions

Features are extracted to find a better fit for the intersection map. These features are obtained by using training maps. An intersection map represents common regions in both maps. The photo and text train maps are the regions in the initial photo and text maps except the intersection regions which are given in Eq. 3.24.

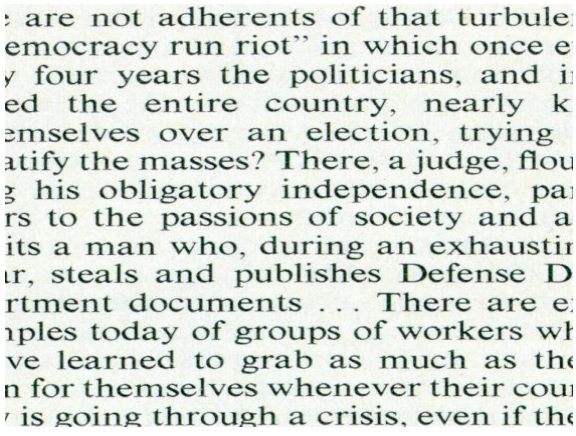
$$
\begin{aligned}
intersectMap &= PhotoMap \cap TextMap \\
PhotoTrainMap &= PhotoMap - intersectMap \\
TextTrainMap &= TextMap - intersectMap.
\end{aligned}
\tag{3.24}
$$

Train maps are computed as shown in Eq. 3.24. The photo train map includes only photo region and the text train map consists of only text region as well. After training and test maps are defined, three different features, standard deviation (SD) in x and y-direction and entropy, are utilized to classify the regions in the intersection map.

3.5.2 Feature extraction

Standard Deviation in x and y direction

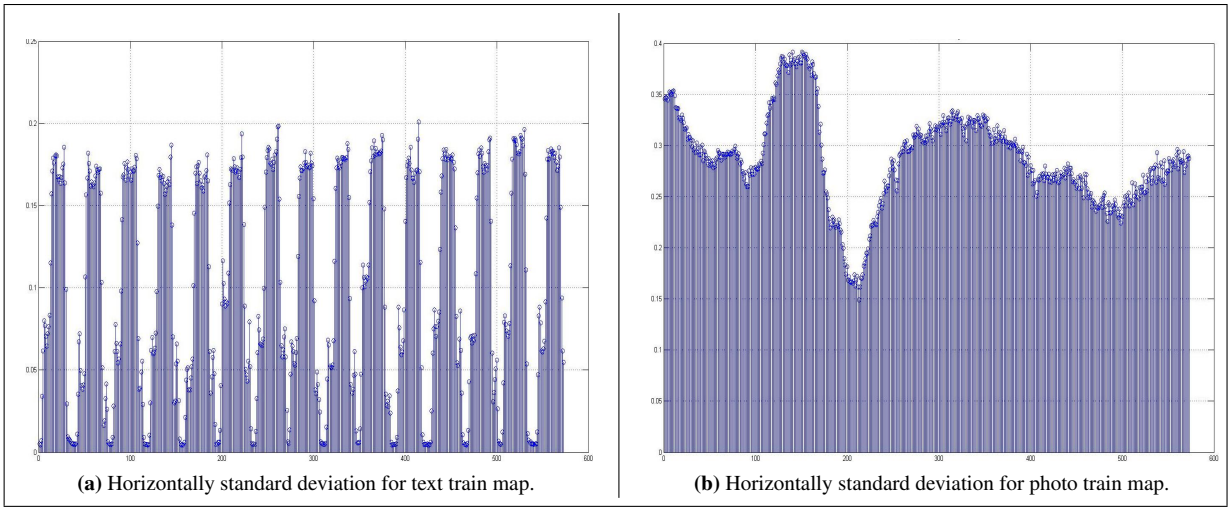
SD in x and y direction are computed by dividing the train maps into the same fixed window size used in block-wise segmentation, $S \times S$. Each window gives one SD value which constitutes a vector for the entire region in both train maps. To illustrate the concept, train maps for both regions and plots of the corresponding features are presented in Fig. 3.18.



(a) Text train map.

(b) Photo train map.

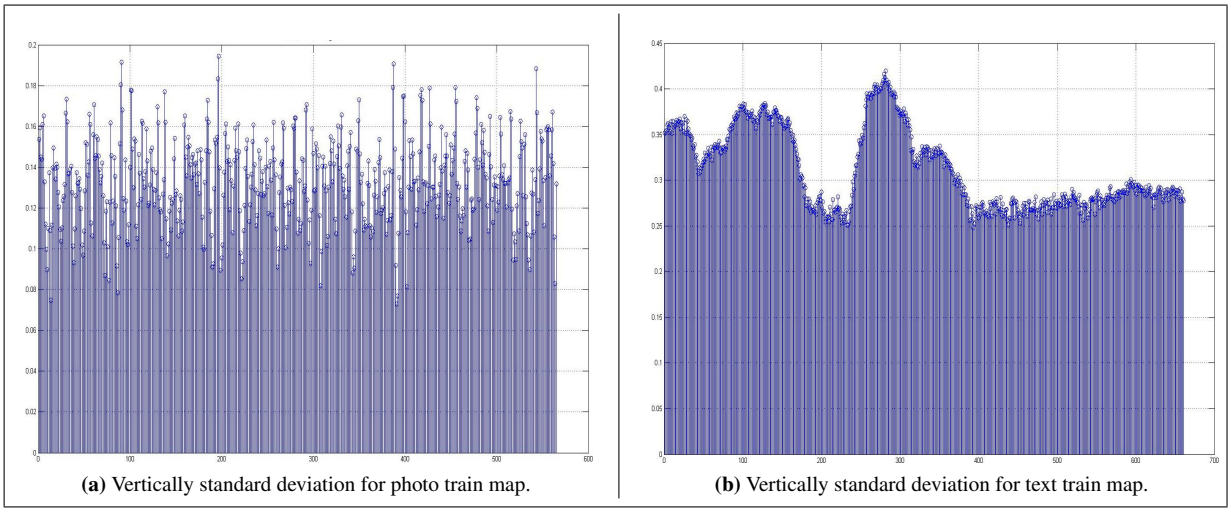
Fig. 3.18: Train maps.



(a) Horizontally standard deviation for text train map.

(b) Horizontally standard deviation for photo train map.

Fig. 3.19: Standard deviation of the train maps in horizontal direction.



(a) Vertically standard deviation for photo train map.

(b) Vertically standard deviation for text train map.

Fig. 3.20: Standard deviation of the train maps in vertical direction.

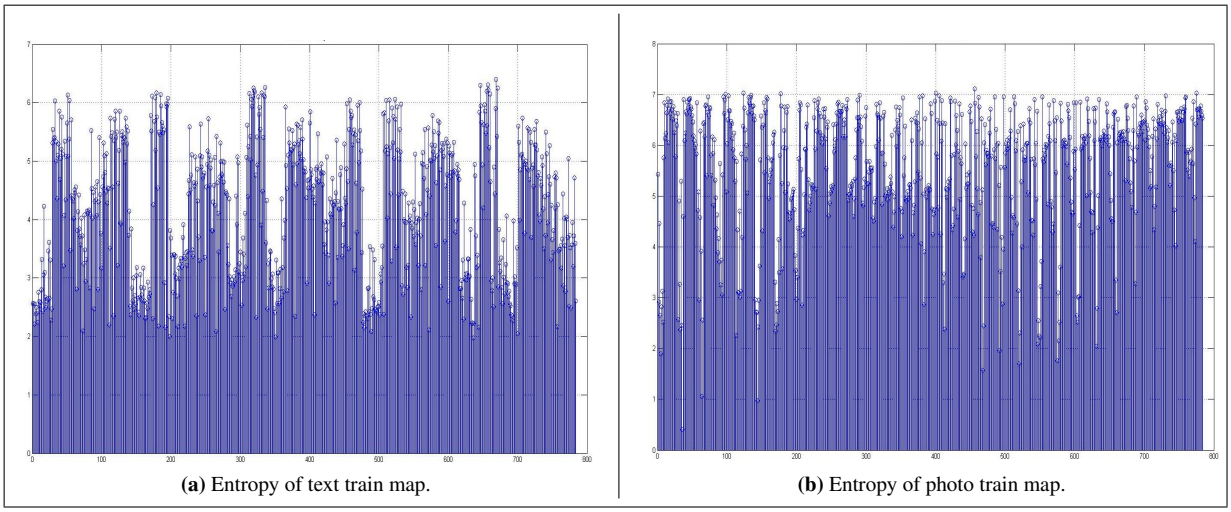
Not unexpectedly, text region have greater standard deviation values in x direction than image region because background and text pixels form a relatively better contrast than the image regions. SD in y direction is also considered because text might be written in y direction as well. However, this case is not applicable for this image. Although the image regions may also form some contrast with the background, it cannot be as high as the text regions particularly in the image region itself, as observed in Fig. 3.19 and 3.20). In addition, an image region can also consist of monotonic pattern or texture which presumes lower standard deviation values. Therefore, greater standard deviation is anticipated to see for the text region both horizontally (x) and vertically (y).

Entropy

Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. It is defined as follows;

$$Ent_R = -\sum_{i=1}^{S}\sum_{j=1}^{S} plog_2(p), \tag{3.25}$$

where p is a vector which contains the probabilities of each gray level that appears in the input image and R is the target region, text or photo. p vector can be easily obtained by utilizing the image histogram. For photo regions, this randomness should be lower than text regions since some areas in photo regions have specific type of pattern like in Fig. 3.21. That is, it is expected to repeat itself in all regions, and result in very low entropy values. However, the text regions have no specific type of pattern unlike photo regions except the spaces between the words. For this reason, they have greater entropy compared to text regions as observed in Fig. 3.21.



(a) Entropy of text train map.

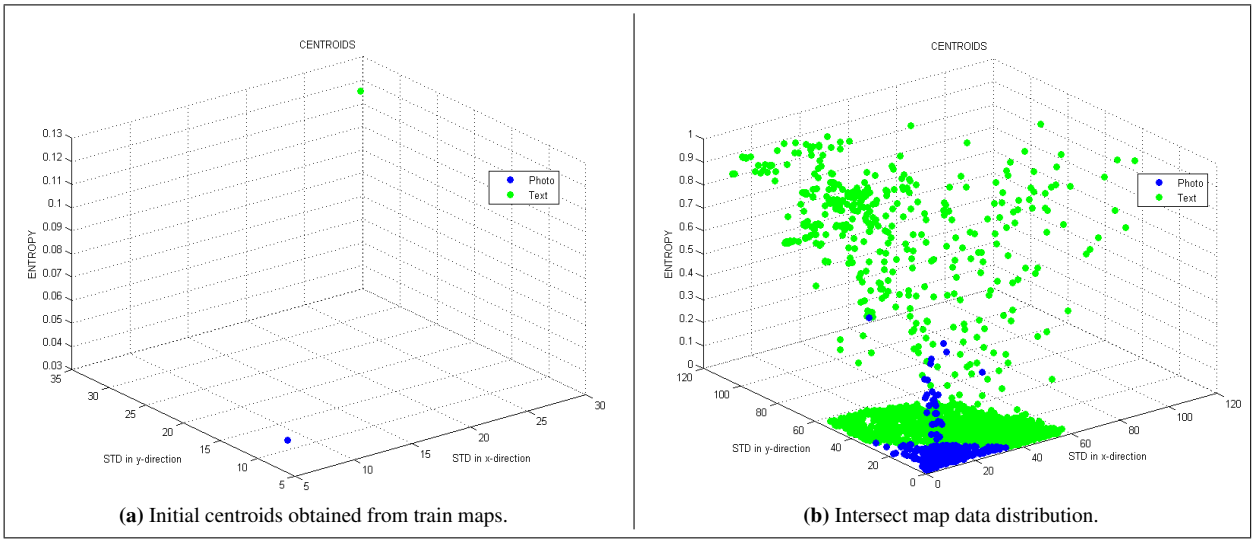
(b) Entropy of photo train map.

Fig. 3.21: Entropy of the train maps.

3.5.3 K-Means algorithm minimizing Euclidean distance

This sub-module uses the intersection map generated in Section 3.5.1. K-Means algorithm is employed to obtain final segmentation map by utilizing the features generated in Section 3.5.2.

Three different features form 3-D space and text, and photo train maps generate two different, separable centroids as shown in Fig. 3.22. The coordinates of the centroids are the average value of SD in x, y direction and the entropy. After several iterations, each block in the intersection map is classified as either a text or an photo region and adds to the corresponding map. However, the some of the class members are mis-classified since the classes are not linearly separable as shown in Fig. 3.22(b). To reduce the computation time, centroids are not updated unlike usual K-Means, since the amount of training data used for extracting features is much larger than the test data.



(a) Initial centroids obtained from train maps.

(b) Intersect map data distribution.

Fig. 3.22: K-Means.

CHAPTER 4: RESULTS AND DISCUSSIONS

The proposed algorithm is tested on a large database that contains a variety of simple to complex color and grey-scale documents. While selecting a database to evaluate the proposed system, there are several quality constraints that are determined in order to test an algorithm in wide aspect. Antonacopoulos summarizes the characteristics that a data-set is required to have to test and examine a performance in detail [60]. The three main desirable characteristics are the followings:

i. *Realistic*: The data-set should span the real documents that are scanned in daily-life at working places.

ii. *Comprehensive*: It should contain various type of documents in order to evaluate the performance and robustness of a proposed algorithm.

iii. *Flexible structured*: It should be easy to find a document in the database if the user needs to pick several documents under a specific condition.

Thus, the MediaTeam document database from Oulu University [61] provided by MediaTeam research group is used to validate the performance of the proposed document classification technique. The test data-set includes many different type of scanned documents such as articles, advertisements, newsletters, business cards or dictionary documents. Among 19 different type of scanned documents, MUSIC, PROGRAM-LISTING and LINE-DRAWING type of the documents are excluded since the figures/images in the documents are considered as line art rather than photos. The generated page layout classification map outlines (as a rectangular box) text and photo regions while detected lines and edges are shown as in a edge map [see Fig. 4.1]. The average execution time for images with average size of 3000×2000 pixels is 15 seconds running on a 2.4 GHz dual core PC implemented in MATLAB®.

The proposed technique is able to produce both pixel-wise and box-wise maps. The box-wise map is generated to measure the accuracy of the technique since the database provides the classification results with bounding rectangles for region representation. The box-wise classification map for each type of scanned document are compared with the ground-truth, exhibited in figures below. The accuracy rates are also presented in tables. Moreover, the results are bench-marked quantitatively and discussed extensively by utilizing confusion matrix (CM) [for detailed explanation of a confusion matrix, see Section 4.1.1]. However, strong edge/line detection results cannot be shown quantitatively in performance evaluation section because their ground-truths are not provided in the MediaTeam Oulu document database. Therefore, line and strong edges classification in pixel-wise and box-wise maps are only demonstrated in Fig. 4.1 below.

Fig. 4.1 illustrates the generated page layout maps for two documents where text, lines, and strong edges are found. The original color documents are shown in Fig. 4.1(a).

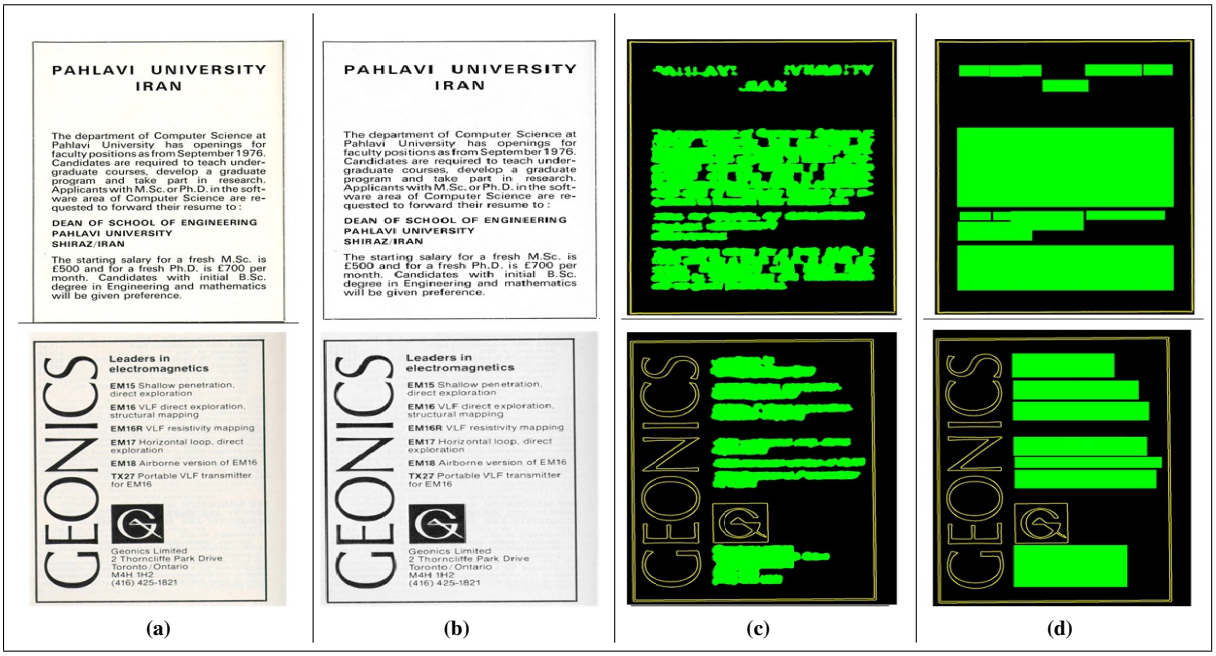


Fig. 4.1: Results for line detection:(a) Original image, (b) enhanced L channel, (c) pixel- and (d) box-wise final classification map.

The color space conversion (RGB to CIEL*a*b*), employed in the proposed algorithm, eliminates artifacts in background regions as shown in Fig. 4.1(b). The enhanced document enables better

detection accuracy as demonstrated in Fig. 4.1(c) and (d) where strong edge/line and text regions are colored in yellow and green, respectively. The documents shown in Fig. 4.1(a) have frames (box-lines) that outline the pages and they are detected fairly accurately in both images. It is worth noticing that the written text with large font-size is detected as strong edges that can be observed in the second example page (second row of Fig. 4.1(a)). Additionally, the pictorial structure shown in the document is also detected. Besides this, note that if the spaces between words are significantly noticeable, the pixel-wise classification can detect these spaces and exclude from the detected text region as can be observed in Fig. 4.1(c). On the other hand, it is not possible to achieve to exclude these spaces from the text region in box-wise segmentation map [see Fig. 4.1(d)] since the boxes are generated according to the coordinates of four corners of a rectangular-box drawn around the classified region.

In Fig. 4.2, three different sample documents are demonstrated to show how the strong edge/line detection algorithm works on the scanned documents that have both photo and text zones. Frames of the first document [see Fig. 4.2(a)] are well-extracted with the text regions. Notice that, although there is no actual line around the photo region in Fig. 4.2(b) and (c), the strong edge/line detection module classifies some pixels as edges. The reason is that they locate at the boundaries so that they are considered as strong edges. However, there are some false detections such as some pixels in the photo region are mis-classified as strong edges in Fig. 4.2(b).

In Fig. 4.3, 4.4, and 4.5, 15 different types of scanned documents whose classes are ADDRESS-LIST, ARTICLE, ADVERTISEMENT, BUSINESS-CARD, CHECK, COLOR SEGMENTATION, CORRESPONDENCE, DICTIONARY, FORM, MANUAL, NEWSLETTER, OUTLINE, PHONE-BOOK, STREET-MAP and TERRAIN-MAP, respectively are illustrated. Photo regions are represented with blue and text zones are shown as green in the ground-truth, colored and grey-scale classification maps. Moreover, cyan regions correspond to a common zone when photo and text regions overlap.

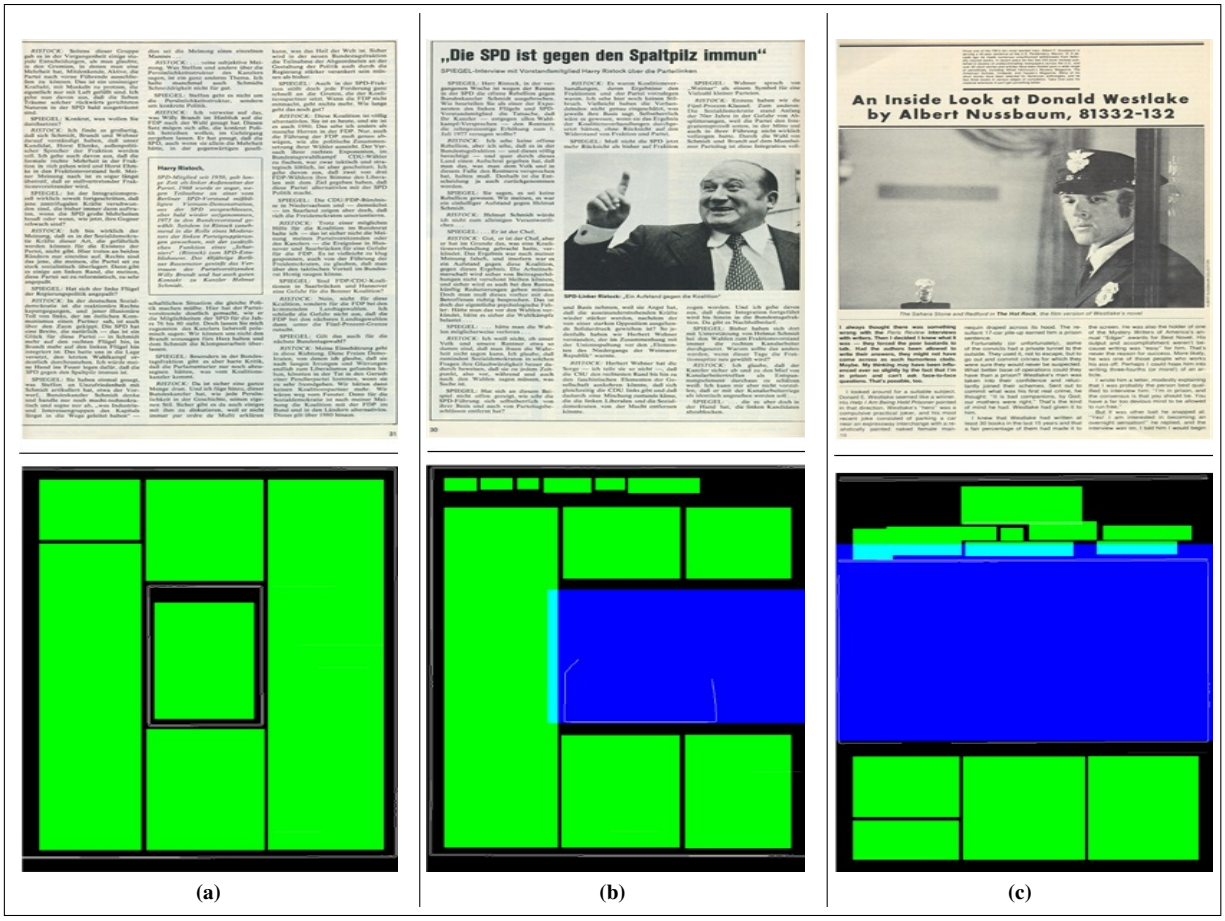


Fig. 4.2: Results for line detection: Document (a) without image, (b) and (c) with strong edge/line, text and photo.

Firstly, a document from ADDRESS-LIST class, whose background can be considered as complex, is shown in Fig. 4.3(a) below. Corresponding color and grey-scale generated page layout classification maps correlate well with the human-generated ground-truth except a tiny region (false positive) in the photo region detected as both photo and text. Fig. 4.3(b) shows an ADVERTISEMENT type of document. The proposed algorithm generates an accurate classification map for both colored and grey documents except for missing a big font-size text photo region (ELEKTOR) under the small photo region at the right bottom. The figure captions (at the top and bottom of the page) in the document of ARTICLE class are mis-classified as background in both classification maps [see Fig. 4.3(c)]. It seems that the algorithm in Fig. 4.3(d) misses the great portion of the photo region compared to the ground-truth although the main body of the region is well-detected.

Fig. 4.3: Final classifications map for:(a) ADDRESS-LIST, (b) ADVERTISEMENT, (c) ARTICLE, (d) BUSINESS-CARD and (e) CHECK scanned document.

However, the missing region contains only three stripes. These stripes introduces false positives since they are connected to photo region and represented in boxes. Moreover, although the text regions in class of BUSINESS-CARD and CHECK documents are separated with significant spaces, the ground-truth classifies these regions in the same box. However, the proposed technique is able to notice these spaces and omit these spaces from the classification map [see Fig. 4.3(d) and (e)].

First of all, it is worth noticing that the colored and grey page layout classification maps for five different types of scanned documents illustrated in Fig. 4.4 are nearly same. The scanned document in Fig. 4.4(a) consists of only photo region. In other words, there is no text region as shown in the classification maps and ground-truth. The CORRESPONDENCE type of document is presented in Fig. 4.4(b). The proposed algorithm manages to detect all text and photo regions.

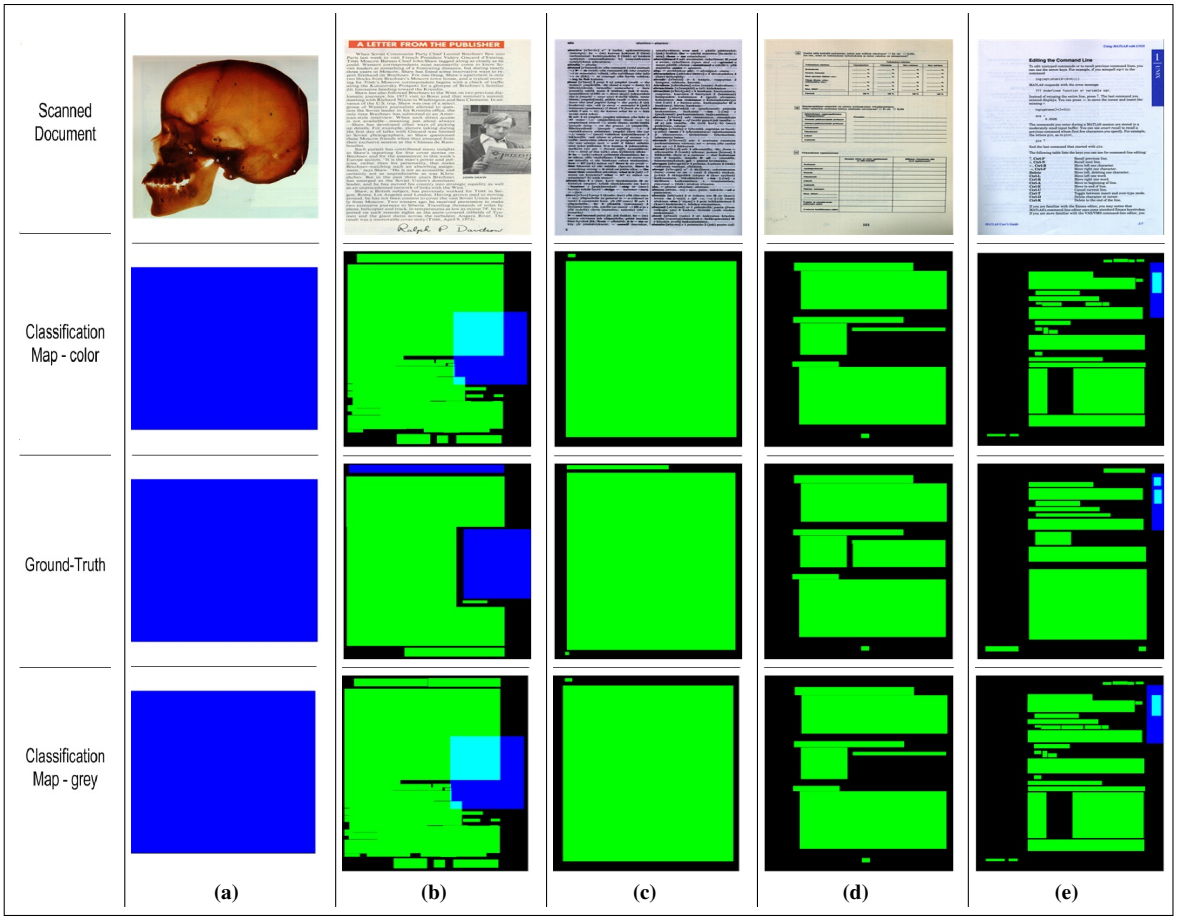


Fig. 4.4: Final classification map for: (a) COLOR SEGMENTATION, (b) CORRESPONDENCE, (c) DICTIONARY, (d) FORM and (e) MANUAL scanned document.

However, some part of the photo region is also classified as text since the text region at the left of the photo is segmented in the same box with top of the photo. Text and photo regions in the DICTIONARY and FORM types of scanned documents are detected fairly accurately except the text regions at the footer and header compared to ground-truth because the database is classified the regions according to their contents [see Fig. 4.4(c), (d), and (e)]. The same phenomena in the Fig. 4.4(e) can be observed at the body of the document in the last text box.

Last set of scanned documents which consists of NEWSLETTER, OUTLINE, PHONE-BOOK, STREET-MAP, and TERRAIN-MAP are presented in Fig. 4.5. Again, it is worth noticing that both page layout classification maps are nearly same. Text zones in NEWSLETTER document are well extracted but photo region at the top of the page (circle with stars) is missed since the main body

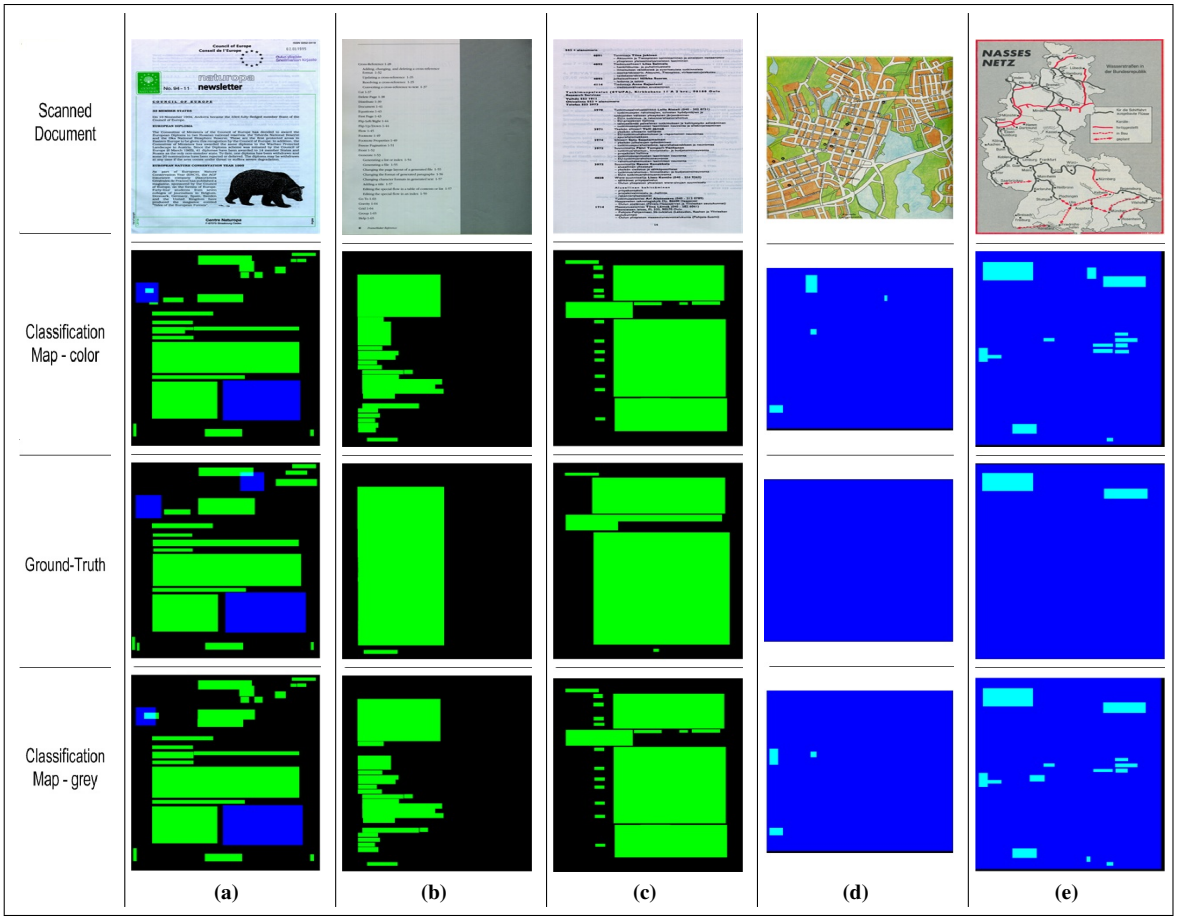
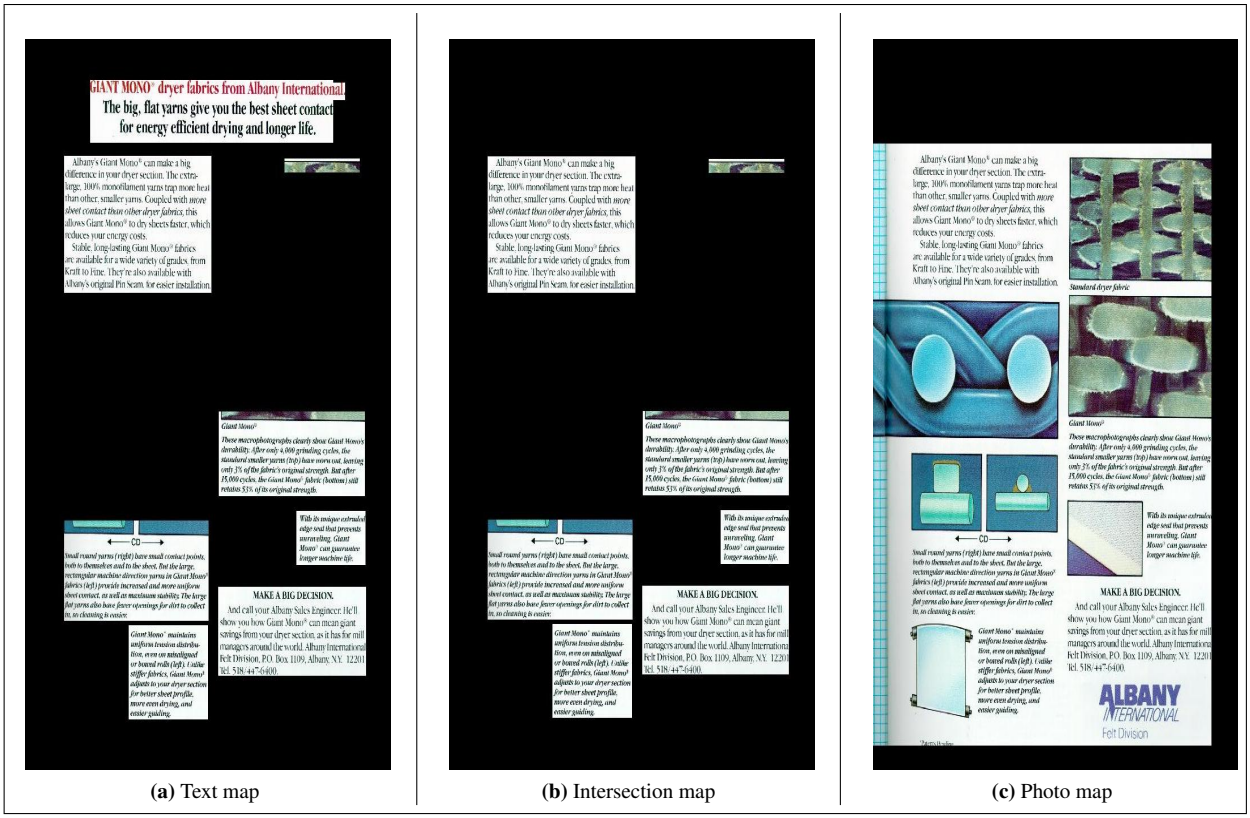


Fig. 4.5: Final classification map for: (a) NEWSLETTER, (b) OUTLINE, (c) PHONE-BOOK, (d) STREET-MAP and (e) TERRAIN-MAP scanned document.

of the region includes background [see Fig. 4.5(a)]. The body of text zone in OUTLINE scanned document in Fig. 4.5(b) is very accurately detected. And notice that, photo module achieves to differentiate a real photo region and the grey part of the background at the left-side of the document which is considered as complex background. Text regions in Fig. 4.5(c) are well-segmented and photo detection module in both segmentation maps is accurate in differentiating the complex background and any photo region. In STREET- and TERRAIN-MAP scanned documents, the maps are classified as photos in the ground-truth and they are well-segmented as photo in segmentation maps, as well [see Fig. 4.5(d) and (e)]. Small text regions, corresponded to country or street names, are mis-classified as text in our classification maps as opposed to the ground-truths.



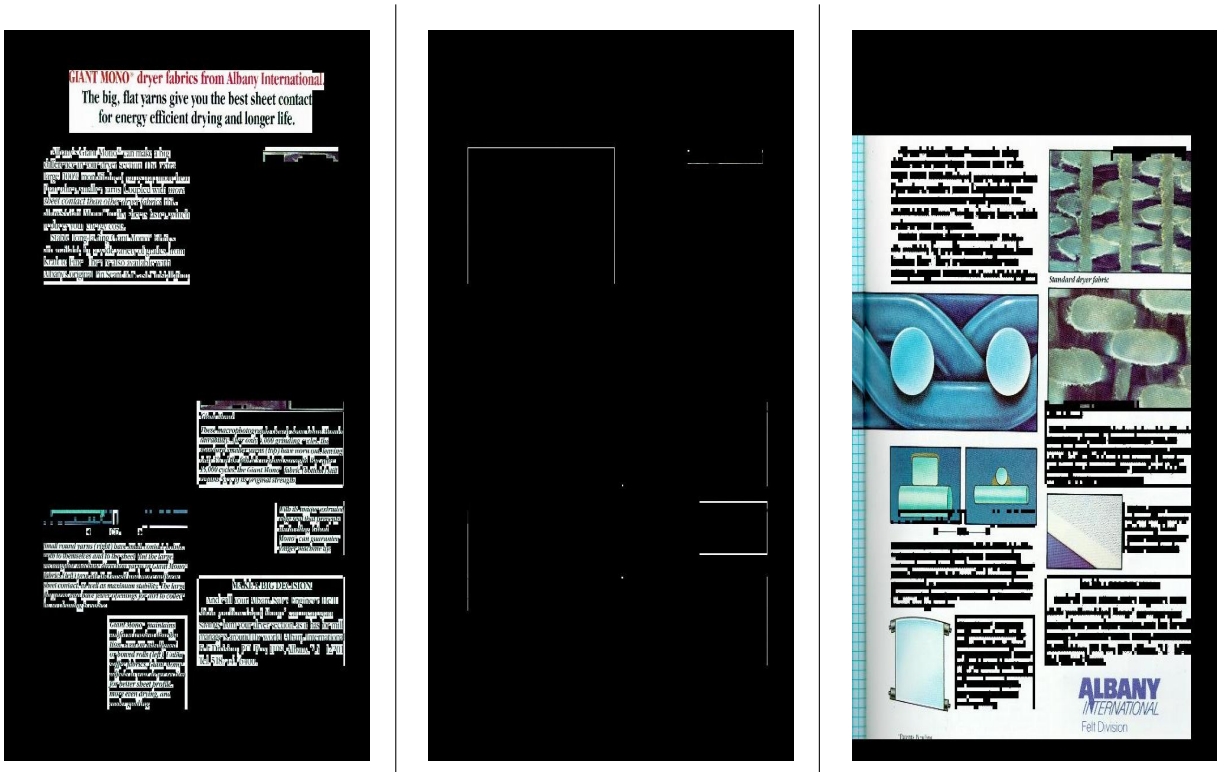
(a) Text map **(b)** Intersection map **(c)** Photo map

Fig. 4.6: Before map combination module.

A demonstration is illustrated in Fig. 4.6 and 4.7 with one of the scanned documents in the database in order to show the obtained results. This module is not included in performance evaluation since the database does not limit the classification maps such that the region of interest (ROI) has to belong either text or photo map. It can be associated with the both segmentation maps. The module is developed for the internal data-set. In Fig. 4.6 above, initial maps obtained from text and photo detection module are presented with the intersection map in Fig. 4.6(b). There are significant portion of text regions detected by the photo detection module shown in Fig. 4.6(c). Besides this, some photo regions are included in the text map which are supposed to be shown only in the photo map [see Fig. 4.6(a)].

Final maps are presented in Fig. 4.7. Text and photo regions are fairly detected and joined to the corresponding correct maps as shown in Fig. 4.7(a) and (c). Nearly all the text regions are included in the final text map by being independent from their font-size although the text train map

does not contain significant information. In addition, the photo regions in the intersection map are well-segmented and joined to the final photo map as well. Final intersection map which contains unclassified pixels/blocks is also shown in Fig. 4.7(b). The experiment results indicate that these pixels/blocks correspond to border of the segmented regions so that they can be represented in either photo or text final map without causing any false positives.



(a) Text map **(b)** Intersection map **(c)** Photo map

Fig. 4.7: After map combination module.

As discussed above, the algorithm is evaluated on a publicly available database of scanned documents and reported the performance in Section 4.1. Furthermore, it has been also tested on various internal document-sets that are gathered from several scanners. The proposed page layout classifier shows enhanced performance independently from scanning technique.

4.1 Performance Analysis and Evaluation

In the past two decades, the developments and improvements in page segmentation techniques to handle with several problems have induced a necessity in performance evaluation. Therefore, various methodologies are introduced in order to evaluate proposed systems. Sufficient performance metrics which are proposed by those methodologies are required to test and evaluate the proposed algorithms comprehensively.

To objectively measure and evaluate the quality of the segmentation results, a feasible and convenient visualization tool is required. In this regard, the Confusion Matrix (CM), introduced by Kohavi & Provost [62] in 1998, is utilized. It is a performance evaluation technique which contains information about actual and predicted classifications obtained by Townsend [63].

4.1.1 Confusion matrix (CM)

In the field of artificial intelligence, a confusion matrix, called also matching matrix, is a way of visualizing a performance of an algorithm used in supervised and unsupervised learning systems. It is a square matrix that represents the count of a classification function's predictions with respect to the actual classifications [63]. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Besides this, each row of the matrix should be added up to 100% in order to have a consistent performance metric, assuming that the actual data (ground-truth) is placed to the rows of the given table. An example for 2×2 confusion matrix is illustrated in Table 4.1 below.

Table 4.1: Sample confusion matrix.

Actual \ Predicted	Negative	Positive
Negative	a	b
Positive	c	d

where

- *a* is the number of correct predictions that an instance is *negative*,

- *b* is the number of incorrect predictions that an instance is *positive*,

- *c* is the number of incorrect of predictions that an instance *negative*, and

- *d* d is the number of correct predictions that an instance is *positive*.

Aim of the technique used in the application is to maximize diagonal entries of the matrix when the performance is evaluated by the confusion matrix. One benefit of using a confusion matrix is that it enables to be observed which classes are labeled accurately and mis-labeled instead of presenting only correct classified units. This yields more comprehensive interpretations about the proposed technique such as where the algorithm fails.

Moreover, several standard measures have been defined for the 2-class confusion matrix.

- The *accuracy* (AC) is the proportion of the total number of predictions that are correctly detected as calculated using Eq. 4.1:

$$AC = \frac{a+d}{a+b+c+d}. \tag{4.1}$$

- The *recall* or *true positive* (TP) rate is the proportion of positive cases that are detected accurately and the false positive (FP) rate is the proportion of negatives cases that are mis-detected as positive, as measured using Eq. 4.2 below:

$$TP = \frac{d}{c+d}, \qquad \text{and} \qquad FP = \frac{b}{a+b}. \tag{4.2}$$

- The *true negative* (TN) rate is defined as the proportion of negatives cases that are determined correctly, while the *false negative* (FN) rate is the proportion of positives cases that are misclassified as negative, formulated in Eq. 4.3 below:

$$TN = \frac{a}{a+b}, \qquad \text{and} \qquad FN = \frac{c}{c+d}. \tag{4.3}$$

- Finally, *precision* (P) is the proportion of the predicted positive cases that are correct, as given in Eq. 4.4 below:

$$P = \frac{d}{b+d}. \tag{4.4}$$

Notice that $TP + FN$ and $TN + FP$ are both equal to 1.

In some cases, the accuracy determined using Eq. 4.1 is not an adequate performance metric when the number of negative cases is much greater than the number of positive cases. For this reason, another performance measure, called *F-Measure*, is introduced by Lewis & Gale [64] as defined in Eq. 4.5.

$$F = \frac{(\beta^2 + 1) * P * TP}{\beta^2 * P + TP}, \tag{4.5}$$

where β has a value from 0 to infinity and is used to control the weight assigned to TP and P.

4.1.2 Quantitative evaluation of the proposed classification technique on different type of scanned documents

The confusion matrices for classification accuracy rates are presented with tables and their corresponding classification maps. Confusion matrices are given separately for ARTICLE, NEWSLETTER, CORRESPONDENCE, and ADVERTISEMENT scanned documents since the data-set is large enough to validate the algorithm. However, MANUAL, OUTLINE and DICTIONARY documents are combined to obtain a larger data-set. Additionally, the classification results for ADDRESS-LIST, PHONE-BOOK, TERRAIN- and STREET-MAP, BUSINESS-CARDS, CHECK, FORM and COLOR SEGMENTATION documents are presented in one confusion matrix.

Article

The proposed algorithm is tested on 233 ARTICLE documents. Three typical article documents are presented below [see Fig. 4.8]. The photo regions in the first and second documents are well-extracted. Main body of the text regions in all three documents are also classified correctly. Notice

that, although the background is not white, the photo and text detection module manage to classify the regions in Fig. 4.8(c). However, the algorithm fails to detect the figure captions as discussed above. The accuracy rate for text, photo and background regions are presented in Table 4.2.

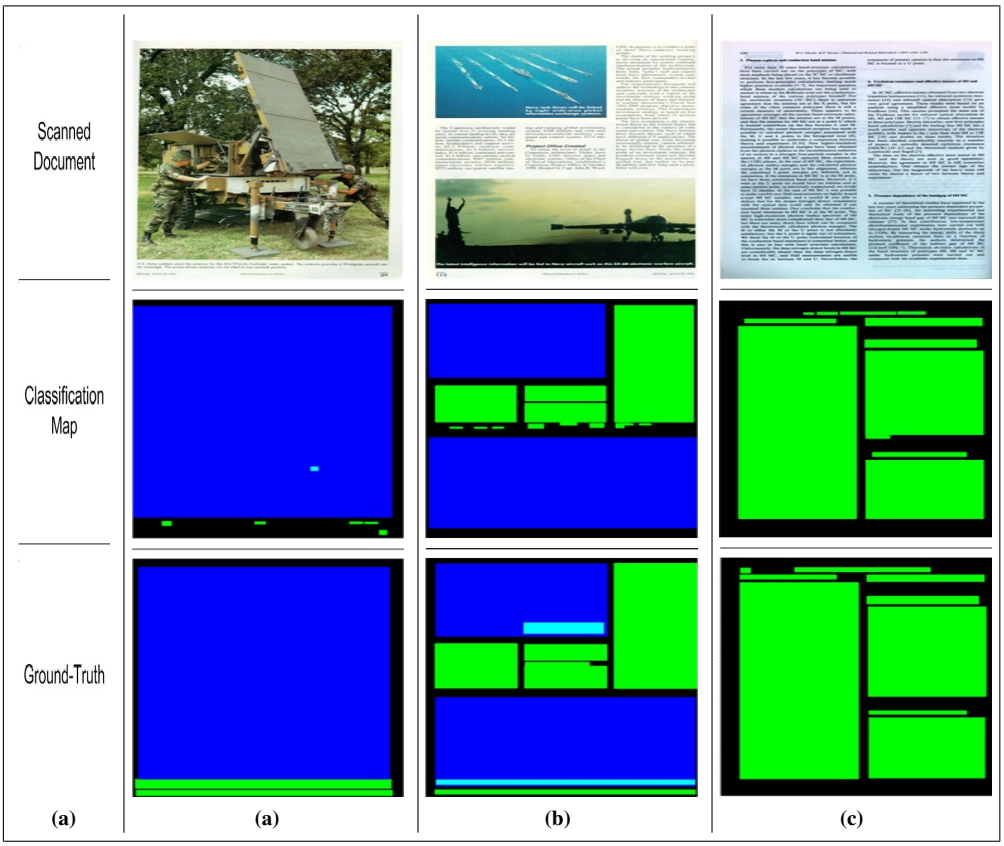


Fig. 4.8: Final classification map for three ARTICLE documents.

Table 4.2: Confusion matrix for ARTICLE document

Ground Truth \ Proposed Algorithm	Text	Photo	Background
Text	0.88	0.03	0.09
Photo	0	0.96	0.04
Background	0.01	0.02	0.97

Newsletter

The algorithm is tested on 42 NEWSLETTER documents. They are similar to the article documents in terms of background structure. Mainly, the background is white and they consist of text

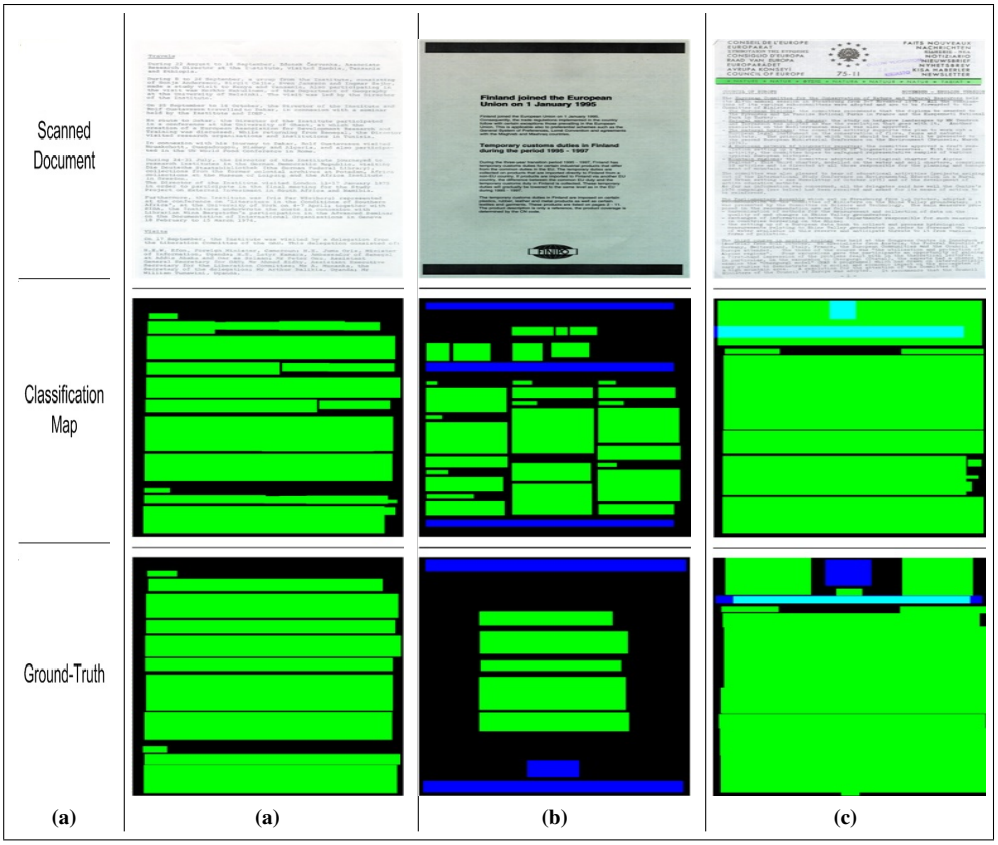


Fig. 4.9: Final classification map for three NEWSLETTER documents.

regions. The text and photo regions are classified well except a few false detections [see Fig. 4.9]. These false detections occur because the classified regions are represented in boxes. Suppose that a line includes a long sentence and it is connected with the previous line which has a short sentence or a word. When the algorithm detects these lines, it draws a box around the long sentence which introduces false detections for the previous line. Table 4.3 summarizes the classification accuracies for the regions.

Table 4.3: Confusion matrix for NEWSLETTER document

Ground Truth \ Proposed Algorithm	Text	Photo	Background
Text	0.89	0.04	0.07
Photo	0.05	0.87	0.08
Background	0.03	0.13	0.84

Correspondence

24 CORRESPONDENCE documents are used for testing purpose. Three different type of correspondence documents are illustrated in Fig. 4.10. The regions are classified accurately except a small photo region in Fig. 4.10(a). Additionally, it is worth noticing that although the background in Fig. 4.10(a) is very complex, photo detection module does not fail in these type of complex background documents.

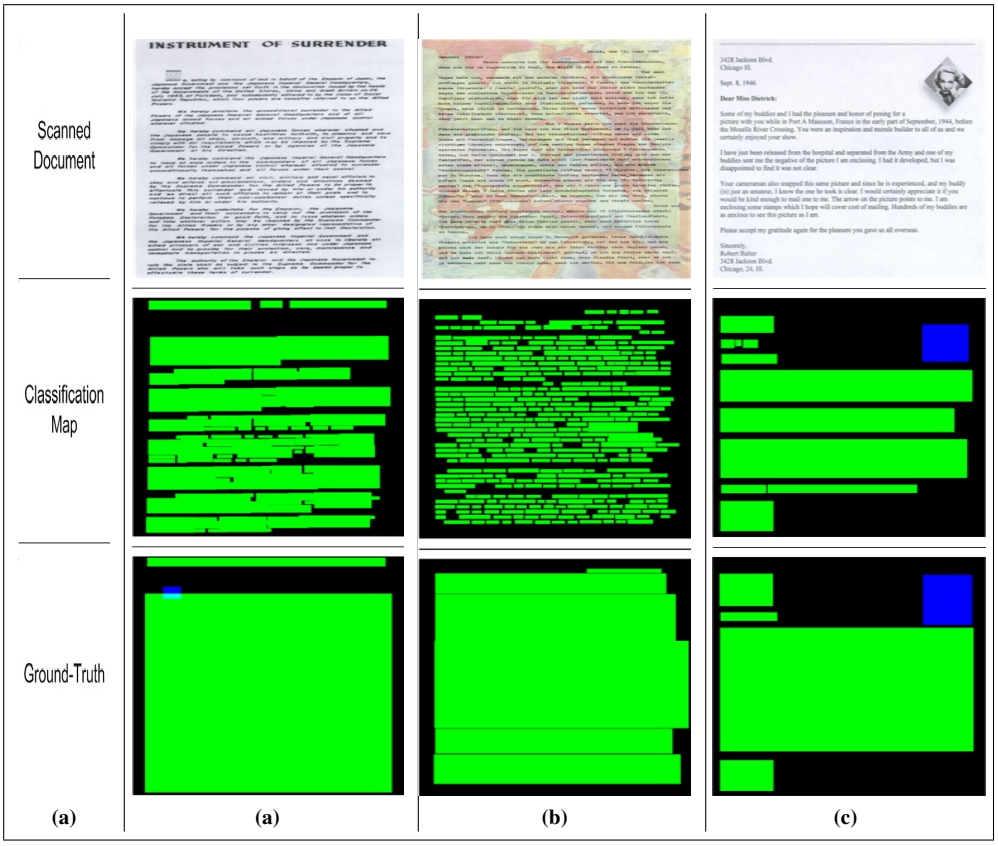


Fig. 4.10: Final classification map for three CORRESPONDENCE documents.

The ground-truth classifies the text regions depending on the content (title, body, header, footer *etc.*). For this reason, it might box the regions including the background [see Fig. 4.10(c)]. On the other hand, the proposed algorithm classifies each region independently so that our boxes mainly do not include any background regions between the paragraphs or words if they are well-separated. Therefore, this phenomena causes false negatives between background and text zones as presented

in Table 4.4 as well.

Table 4.4: Confusion matrix for CORRESPONDENCE document

Ground Truth \ Proposed Algorithm	Text	Photo	Background
Text	0.91	0.01	0.08
Photo	0.02	0.94	0.04
Background	0.01	0.05	0.94

Advertisement

24 ADVERTISEMENT documents are tested to validate the performance of the proposed algorithm. They are considered as complex color documents since they consist a great number of different color tones. The photo and text regions in Fig. 4.11(a) and (b) are classified fairly well. Again, complex background is separated from the real content of the document and the regions are well-segmented [see Fig. 4.11(c)].

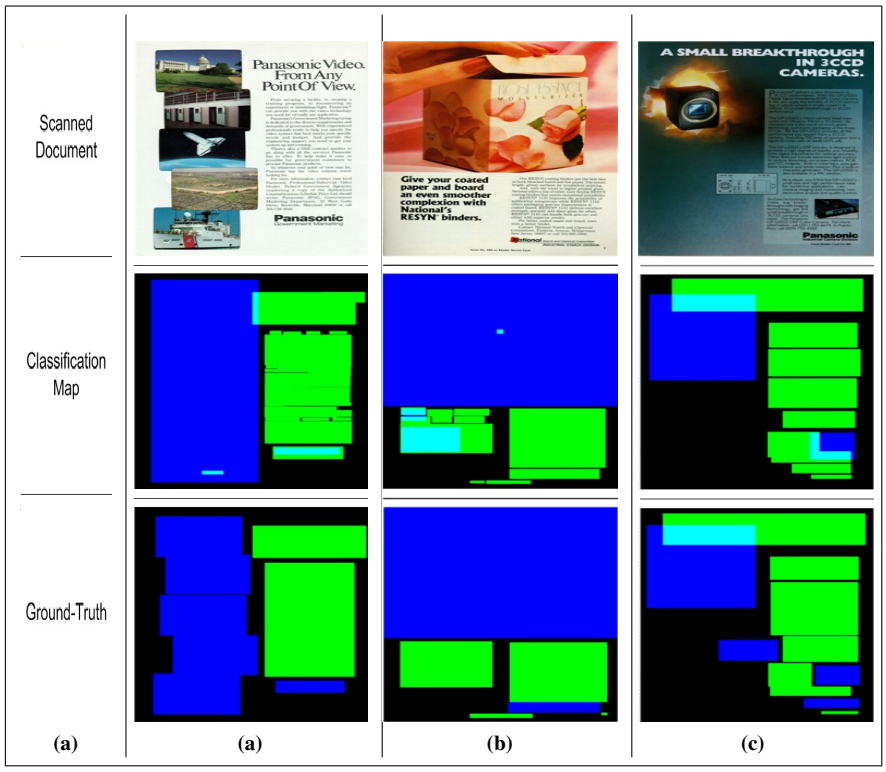


Fig. 4.11: Final classification map for three ADVERTISEMENT documents.

However, the photo regions at the bottom of all documents demonstrated in Fig. 4.11, (PANASONIC and National), are mis-classified as text. This results in false positives between text and photo regions as observed in Table 4.5. Besides this, the algorithm misses the photo region that has same color tone with the background [see Fig. 4.11(c)].

Table 4.5: Confusion matrix for ADVERTISEMENT document

Ground Truth \ Proposed Algorithm	Text	Photo	Background
Text	0.91	0.06	0.03
Photo	0.09	0.89	0.02
Background	0.04	0.12	0.84

Manual & Outline & Dictionary (MOD)

35 manual, 19 outline, and 12 dictionary documents are used to evaluate the performance of the proposed algorithm. One sample document from each class is presented in Fig. 4.12.

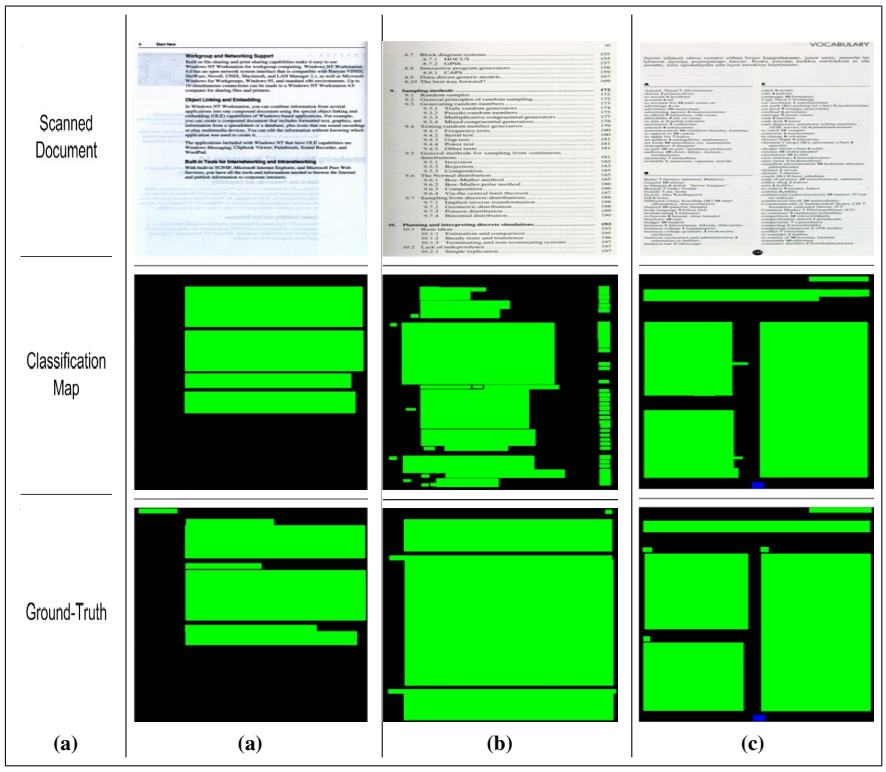


Fig. 4.12: Final classification map for MOD documents.

The reflection at the background is successfully eliminated from the original document and text regions are classified correctly except the footer at the top left corner [see Fig. 4.12(a)]. In Fig. 4.12(b) and (c), text regions are detected with very high accuracy. A few regions causes false negatives because of representing the regions in box-wise. Again, our algorithm is able to separate the text regions if they are disconnected enough. Nevertheless, the ground-truth classifies the regions according to their contents as mentioned above.

Table 4.6: Confusion matrix for MOD documents

Ground Truth \ Proposed Algorithm	Text	Photo	Background
Text	0.87	0.03	0.1
Photo	0	0.92	0.08
Background	0	0.09	0.91

Address-list, Phone-book, Terrain- & Street-map, Business-cards, Check, Form and Color segmentation (OTHER)

Totally 62 documents are utilized in this section. One scanned document from each class is demonstrated to demonstrate the results in Fig 4.13. Some false negatives are observed in Fig. 4.13(a) and (b) because of high-frequency content in the document. Although there are some text regions in Fig. 4.13(b), the ground-truth does not consider them as text. The classification maps given in Fig 4.13(c), (d) and (h) are fairly matches with the ground-truth. The text and photo regions are well-detected and classified correctly with few exceptions because of box-wise representation phenomena [see Fig. 4.13(e), (f) and (g)].

Table 4.7: Confusion matrix for OTHER documents

Ground Truth \ Proposed Algorithm	Text	Photo	Background
Text	0.84	0.01	0.15
Photo	0.03	0.79	0.18
Background	0	0.05	0.95

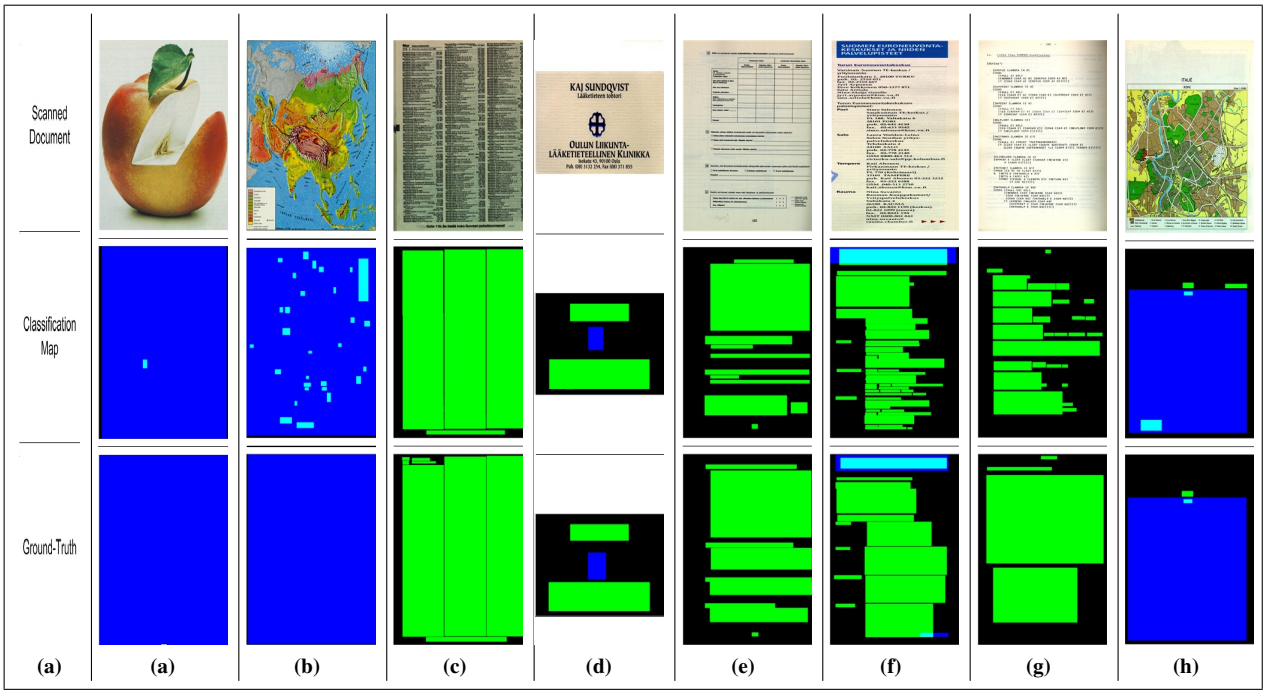


Fig. 4.13: Final classification map for OTHER documents.

4.2 Comparison with the techniques in literature

In this section, the performance of the proposed technique is compared with two algorithms from the state-of-the-art. All of the proposed techniques [26, 39] are used pixel accuracy rate as a performance metric.

4.2.1 Comparison to work done by Duong *et al.*[26]

Duong *et al.* propose a two-step document analysis system which detects the regions using cumulative gradient considerations and classifies the regions as text and non-text zones utilizing geometric and texture features [26]. Entire MediaTeam document database [61] that includes $\sim$ 500 document images, is employed to validate the performance of the proposed system. Text regions are extracted approximately as rectangular areas defined by their bounding boxes. For each scanned document, the accuracy rates are obtained by the following Eq. 4.6:

$$\text{Accuracy}(\%) = \frac{|t|}{|T|}, \tag{4.6}$$

where T is the set of the rectangular text regions defined by the database and t is the set of text regions segmented successfully by the proposed system [26]. According the performance metric given in Eq. 4.6, the accuracy rate comparison for each class between our algorithm and the proposed technique in [26] is presented in Table 4.8.

Table 4.8: Performance comparison between Duong *et al.*[26] and Our classification technique

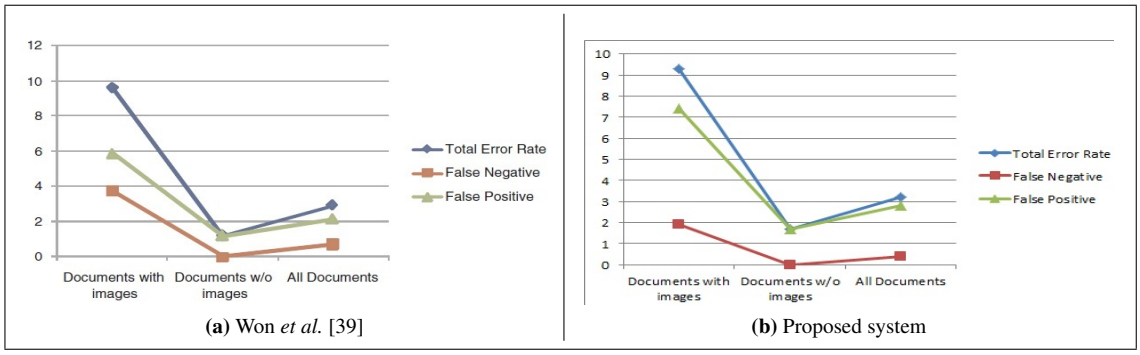
Document Class (MediaTeam labels)	Number of samples	Average Performance (Av. Perf.) [60]	Our Av. Perf.
ADDRESS-LIST	6	0.75	0.81
ADVERTISEMENT	24	0.95	0.91
ARTICLE	233	0.75	0.88
BUSINESS-CARDS	11	0.96	0.91
CHECK	3	0.93	0.81
COLOR-SEG-IMAGES	10	N/A	N/A
CORRESPONDENCE	24	0.82	0.91
DICTIONARY	12	0.97	0.95
FORM	23	0.86	0.82
MANUAL	35	0.88	0.87
NEWSLETTER	42	0.86	0.89
OUTLINE	19	0.84	0.80
PHONE-BOOK	7	0.88	0.93
PROGRAM-LISTING	12	0.92	0.78
STREET-MAP	3	1.00	0.87
TERRAIN-MAP	5	0.93	0.90
MATH	17	0.67	0.78
MUSIC	9	0.84	N/A
LINE-DRAWING	7	0.95	N/A

Since no text zone at COLOR-SEG-IMAGES is specified in their corresponding ground-truth, this class is not considered. The grey-scale version of the documents are tested in [26] although they are scanned in RGB space. However, since our proposed system has no dependency on the number of channels of an input image, the comparison is performed. MUSIC and LINE-DRAWING documents are excluded in our data-set. In most of the documents, both algorithm provide almost same accuracy rates. Our algorithm performs better significantly in ARTICLE,

CORRESPONDENCE and MATH classes as shown in Table4.8. On the contrary, the work in [26] gives considerably better classification results in CHECK and PROGRAM-LISTING. When all the document types are considered, the overall performance of the proposed algorithm achieves 87% accuracy while the methodology in [26] performs with an average of 81% accuracy rate.

4.2.2 Comparison to work done by Won [39]

An algorithm for extracting images in digital documents has been proposed by Won [39] where 233 article scanned documents provided by MediaTeam [61] are utilized to evaluate the algorithm. The proposed technique is also used at the photo detection module in our proposed technique with some modifications. The document is enhanced by pre-processing module which is not applicable in Won's study. In addition, the optimal block size is fixed depending on the size of the scanned document and post-processing technique is utilized instead using a block-size reduction step since they are very computationally expensive process. The performance results are presented in terms of total error rate, false negative and positive which are explained and discussed explicitly in Section 4.1.1. Fig. 4.14 shows the false detections including total error rates.



(a) Won *et al.* [39] **(b)** Proposed system

Fig. 4.14: Error rates(%).

Documents that have photos/images introduce more false positives in our proposed technique compare to the study in [39] since the optimal block-size is fixed and block-size reduction is not applied. Therefore, the pixels at the boundaries are detected as image although they are non-image pixels. However, non-image regions are better classified in our system as seen in false negative

plot. Moreover, false positives in documents without images (i.e.,non-image regions are classified as image regions) are less than the proposed technique in [39]. This implies that pre-processing module in our system provides a document that has better enhanced text and image regions. This prevents our system from a text region to be mis-classified as an image region. Note that, the false negatives in documents without images is zero for both proposed systems since there is no image region defined. The average error rate for all 233 documents in [39] and our technique is $\sim 2.9\%$ and $\sim 4.1\%$, respectively [see Fig. 4.14]. However, more importantly, although his algorithm is developed for grey-scale scanned documents which are assumed to have a white background region, our algorithm achieves 89% accuracy in ~ 500 documents photo classification in both RGB and grey-scaled scanned documents.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

An algorithm for document classification has been proposed in this paper where text, photo, and strong edge/line regions are identified for both color and grey-scale scanned documents. The proposed technique is primarily based on wavelet decomposition, run-length encoding, projection based on basis vectors, MRF-MAP optimization, Hough transform and edge linkages and a merging procedure by using K-Means clustering. It has been tested on large databases of scanned documents, obtained by different scanning techniques. A variety of simple, complex, color, and grey-scale documents are used to evaluate the proposed technique. Experimental results indicate that the algorithm works with an average of 85% accuracy for text, photo and background regions on both color and gray-level scanned documents. And more importantly, it provides consistent results for different types of documents. For this reason, it gives an opportunity to use on several different types of scanned documents where the other methods cannot provide this feasibility. Although they usually consider one specific type of scanned document, the proposed technique achieves very close accuracy compared to the other proposed methods as represented with numbers in the results presented above. The accomplishments of the proposed work are summarized below.

1) The proposed algorithm is an efficient classification method designed for colored and grey-scale scanned documents.

2) Text, photo, strong-edge/line and background regions can be shown in one classification map instead of only segmenting the document as text and non-textual region.

3) Scanning artifacts and reflections occurred in the background are eliminated so that the algorithm is a robust technique against these artifacts.

4) High accuracy is achieved although several different types of simple and complex colored scanned documents are employed in testing stage.

5) It is independent from a scanning technique since the algorithm is tested on MediaTeam database and documents obtained from the industry.

The objectives of future work are to:

1) Increase the overall performance and reduce the computation time because it yields more reliable system for document retrieval applications. Some identified drawbacks are:
   - Mis-detection of figure captions or text regions written on photo region.
   - Lower accuracy in lower resolution scanned documents.
2) Recognize additional objects such as gradients, 1D or and 2D bar-codes.
3) Develop a computationally efficient algorithm for text recognition, mainly OCR, from scanned images.
4) Assign semantic and logical relations to the classified text, strong edge/line, photo and background regions by utilizing the generated map.

REFERENCES

[1] L. O'Gorman and R. Kasturi, *Document Image Analysis*. IEEE Computer Society Press, 1997.

[2] R. Kasturi, L. O'Gorman, and V. Govindaraju, "Document image analysis: A primer," *Sadhana*, vol. 27, no. 1, pp. 3–22, 2002.

[3] F. Shih, S. Chen, D. Hung, and P. Ng, "A document segmentation, classification and recognition system," in *Proceedings of the 2nd International Conference on Systems Integration (ICSI)*, 1992, pp. 258–267.

[4] H. Fujisawa, Y. Nakano, and K. Kurino, "Segmentation methods for character recognition: From segmentation to document structure analysis," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1079–1092, 1992.

[5] S. Mao, A. Rosenfeld, and T. Kanungo, "Document structure analysis algorithms: A literature survey," in *Proceedings of SPIE Electronic Imaging*, vol. 5010, 2003, pp. 197–207.

[6] W. Croft and D. Harper, "Using probabilistic models of document retrieval without relevance information," *Journal of documentation*, vol. 35, no. 4, pp. 285–295, 1993.

[7] S. Yu, D. Cai, J. Wen, and W. Ma, "Improving pseudo-relevance feedback in web information retrieval using web page segmentation," in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 11–18.

[8] D. Doermann, "The retrieval of document images: A brief survey," in *Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, 1997, pp. 945–949.

[9] N. Arica and F. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications*

and Reviews, vol. 31, no. 2, pp. 216–233, 2001.

[10] M. Mitra and B. Chaudhuri, "Information retrieval from documents: A survey," *Information Retrieval*, vol. 2, no. 2, pp. 141–163, 2000.

[11] J. Fisher, S. Hinds, and D. D'Amato, "A rule-based system for document image segmentation," in *Proceedings of 10th International Conference on Pattern Recognition*, vol. 1, 1990, pp. 567–572.

[12] F. Esposito, D. Malerba, G. Semeraro, E. Annese, and G. Scafuro, "An experimental page layout recognition system for office document automatic classification: An integrated approach for inductive generalization," in *Proceedings of 10th International Conference on Pattern Recognition*, vol. 1, 1990, pp. 557–562.

[13] R. Haralick, "Document image understanding: Geometric and logical layout," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)*, 1994, pp. 385–390.

[14] A. Zlatopolsky, "Automated document segmentation," *Pattern Recognition Letters*, vol. 15, no. 7, pp. 699–704, 1994.

[15] D. Sharma and B. Kaur, "Document Image Segmentation Using Recursive Top-Down Approach and Region Type Identification," *Information Processing and Management*, vol. 70, pp. 571–576, 2010.

[16] Z. Shi and V. Govindaraju, "Multi-scale techniques for document page segmentation," in *Proceedings of 8th International Conference on ICDAR*, vol. 2, 2005, pp. 1020–1024.

[17] F. Wahl, K. Wong, and R. Casey, "Block segmentation and text extraction in mixed text/image documents," *Computer Graphics and Image Processing*, vol. 20, no. 4, pp. 375–390, 1982.

[18] S. Lam, D. Wang, and S. Srihari, "Reading newspaper text," in *Proceedings of 10th International Conference on Pattern Recognition*, vol. 1, 1990, pp. 703–705.

[19] A. Antonacopoulos and R. Ritchings, "Flexible page segmentation using the background," in *Proceedings of the 12th International on Pattern Recognition: Computer Vision & Image Processing*, vol. 2, 1994, pp. 339–344.

[20] D. Drivas and A. Amin, "Page segmentation and classification utilising a bottom-up approach," in *Proceedings of the 3rd ICDAR*, vol. 2, 1995, pp. 610–614.

[21] A. Aho, J. Hopcroft, and J. Ullman, *Data structures and algorithms*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1983.

[22] A. Simon, J. Pret, and A. Johnson, "A fast algorithm for bottom-up document layout analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 273–277, 1997.

[23] A. Jain and B. Yu, "Document representation and its application to page decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 294–308, 1998.

[24] S. Grover, K. Arora, and S. Mitra, "Text extraction from document images using edge information," in *Annual IEEE India Conference (INDICON)*, 2009, pp. 1–4.

[25] A. Jain and S. Bhattacharjee, "Text segmentation using Gabor filters for automatic document processing," *Machine Vision and Applications*, vol. 5, no. 3, pp. 169–184, 1992.

[26] J. Duong, M. Côte, H. Emptoz, and C. Suen, "Extraction of text areas in printed document images," in *Proceedings of the 2001 ACM Symposium on Document engineering*, 2001, pp. 157–165.

[27] T. Randen and J. Husoy, "Segmentation of text/image documents using texture approaches," in *Proceedings of the NOBIM-konferansen*, 1994, pp. 60–67.

[28] L. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, pp.

910–918, 1988.

[29] K. Tombre, S. Tabbone, L. Pélissier, B. Lamiroy, and P. Dosch, "Text/graphics separation revisited," in *Proceedings of the 5th International Workshop on Document Analysis Systems V*, 2002, pp. 200–211.

[30] M. Lin, J. Tapamo, and B. Ndovie, "A texture-based method for document segmentation and classification," *South African Computer Journal*, vol. 36, pp. 49–56, 2006.

[31] S. Wang and H. Baird, "Feature selection focused within error clusters," in *19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[32] S. Wang, H. Baird, and C. An, "Document content extraction using automatically discovered features," in *10th International Conference on Document Analysis and Recognition*, 2009, pp. 1076–1080.

[33] S. Chaudhury, M. Jindal, and S. Dutta Roy, "Model-guided segmentation and layout labelling of document images using a hierarchical conditional random field," *Pattern Recognition and Machine Intelligence*, vol. 5909, pp. 375–380, 2009.

[34] H. Baird, M. Moll, C. An, and M. Casey, "Document image content inventories," in *Proceedings of SPIE/IS&T Document Recognition & Retrieval XIV Conference*, vol. 6500, 2007.

[35] Y. Zheng, H. Li, and D. Doermann, "Machine printed text and handwriting identification in noisy document images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 337–353, 2004.

[36] C. Shin, D. Doermann, and A. Rosenfeld, "Classification of document pages using structure-based features," *International Journal on Document Analysis and Recognition*, vol. 3, no. 4, pp. 232–247, 2001.

[37] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. Joshi, "Text extraction and document image segmentation using matched wavelets and MRF model," *IEEE Transactions on Image*

Processing, vol. 16, no. 8, pp. 2117–2128, 2007.

[38] L. Caponetti, C. Castiello, and P. Górecki, "Document page segmentation using neuro-fuzzy approach," *Applied Soft Computing*, vol. 8, no. 1, pp. 118–126, 2008.

[39] C. Won, "Image extraction in digital documents," *Journal of Electronic Imaging*, vol. 17, no. 7, pp. 1–7, 2008.

[40] J. Gllavata, R. Ewerth, and B. Freisleben, "Text detection in images based on unsupervised classification of high-frequency wavelet coefficients," vol. 1, 2004, pp. 425–428.

[41] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[42] C. Burrus, R. Gopinath, and H. Guo, *Introduction to wavelets and wavelet transforms: A primer*. Prentice-Hall, Englewood Cliffs, NJ, 1998.

[43] Y. Sheng, *Wavelet transform*. Boca Raton, Fl: CRC, 1996.

[44] A. Calderbank, I. Daubechies, W. Sweldens, and B. Yeo, "Wavelet transforms that map integers to integers* 1," *Applied and computational harmonic analysis*, vol. 5, no. 3, pp. 332–369, 1998.

[45] J. Hammersley and P. Clifford, "Spatial interaction and the statistical analysis of lattice systems," *unpublished*.

[46] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 192–236, 1974.

[47] S. Li, *Markov random field modeling in image analysis*. Springer-Verlag New York Inc, 2009.

[48] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[49] A. Tekalp, *Digital Video Processing*. Prentice Hall Signal Processing Series, 1995.

[50] H. Derin and H. Elliott, "Modeling and segmentation of noisy and textured images using Gibbs random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 1, pp. 39–55, 1987.

[51] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 48, no. 3, pp. 259–302, 1986.

[52] R. Duda and P. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, 1972.

[53] F. Wahl, L. Abele, and W. Scherl, "Merkmafleu er die Segmentation von Dokumenten zur Automatischen Textverarbeitung," in *Proceedings of the 4th DAGM-Symposium Hamburg, Federal Republic of German*, 1981.

[54] L. Abele, F. Wahl, and W. Scherl, "Procedures for an Automatic Segmentation of Text Graphic and Halftone Regions in Documents," in *Proceedings of the 2nd Scandinavian Conference on Image Analysis*, 1981.

[55] K. Wong, R. Casey, and F. Wahl, "Document analysis system," *IBM journal of research and development*, vol. 26, no. 6, pp. 647–656, 1982.

[56] X. Qian and G. Liu, "Text detection, localization and segmentation in compressed videos," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, vol. 2, 2006, p. II.

[57] S. Hinds, J. Fisher, and D. D'Amato, "A document skew detection method using run-length

encoding and the hough transform," in *Proceedings of 10th International Conference on Pattern Recognition*, vol. 1, 1990, pp. 464–468.

[58] P. Green and L. MacDonald, "Colour engineering: Achieving device independent colour," *J. Electron. Imaging*, vol. 13, p. 663, 2004.

[59] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[60] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher, "A realistic dataset for performance evaluation of document layout analysis," in *10th International Conference on Document Analysis and Recognition*, 2009, pp. 296–300.

[61] Sauvola, J. and Kauniskangas, H., "Media team document database II," http://www.mediateam.oulu.fi/downloads/MTDB/, University of Oulu, Finland, 1999.

[62] R. Kohavi and F. Provost, "Editorial for the special issue on applications of machine learning and the knowledge discovery process," *Machine Learning*, vol. 30, no. 1, pp. 2–3, 1998.

[63] J. Townsend, "Theoretical analysis of an alphabetic confusion matrix," *Attention, Perception, & Psychophysics*, vol. 9, no. 1, pp. 40–50, 1971.

[64] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 1994, pp. 3–12.

VITA

M. Sezer was born in Denizli, Turkey, in 1986. He received his B.Sc. degree in Electrical & Electronics Engineering from Koc University, Istanbul, in 2009. Same year, he joined Computer Vision and Image & Video Processing Laboratory of Rochester Institute of Technology (RIT), at the Department of Electrical and Microelectronic Engineering as a M.Sc candidate. Sezer is a student member of IEEE and SPIE.