Document Image Processing - A Review

Shazia Akram
Research Scholar
University of Kashmir (India)

Dr. Mehraj-Ud-Din Dar
Director, IT & SS
University of Kashmir (India)

Aasia Quyoum
Research Scholar
University of Kashmir (India)

ABSTRACT
The field of a digital-image processing has experienced dramatic growth and increasingly widespread applicability in recent years. Fortunately, advances in computer technology have kept pace with the rapid growth in volume of image data in these and other applications. Digital-image processing has become economical in many fields of research and in industrial and military applications. While each application has requirements unique from the others, all are concerned with faster, cheaper, more accurate, and more extensive computation.

Analysis of document images for information extraction has become very prominent in recent past. Wide variety of information, which has been conventionally stored on paper, is now being converted into electronic form for better storage and intelligent processing. This needs processing of documents using image analysis, processing methods. This article provides an overview of various methods used for digital image processing using three main components: Pre-processing, Feature extraction and the Classification. Pre-processing includes Image acquisition, Binarization, identification, Layout analysis, feature extraction and classification. Classification is an important step in Office Automation, Digital Libraries, and other document image analysis applications. This article examines the various methods used for document image processing in order to achieve a processed document having high quality, accuracy and fast retrieval.

Keywords
Document, Analysis, Processing, Classification

1. INTRODUCTION
Traditionally, our main form of transmission & storage for information has been by paper documents. These documents include many common types: business letters, forms, engineering drawing & maps, text books, technical manual, music notations & other symbolic data. Though paper was the exclusive medium in past, many documents now originate on the computers & often reside exclusively in electronic form. In spite of this it is unclear whether the computer has decreased/increased the amount of paper document produced, as these are printed out for reading, dissemination, markup predictions of paperless offices made so frequently during the early 1980 has given way to the realization that the objective is not elimination of paper but the ability to deal with the flow of electronic & paper document in effective & integrated way. Document processing in any organization whether having its operations manual or computerized, forms an essential activity in its functioning's. Within document processing, the key activity prior to all other activities is the recognition of documents and hence their categorization [1] [7] [10]. Several good solutions exist for document processing and analysis, this paper tries to give general idea for document processing and the various steps/methods used for that. This will give an overview for processing, analysis and classification of document images

2. DOCUMENT IMAGE ANALYSIS
The objective of Document Image analysis is to recognize the text & graphics components in image of documents & to extract intended information from them. Two categories of document image analysis can be defined.

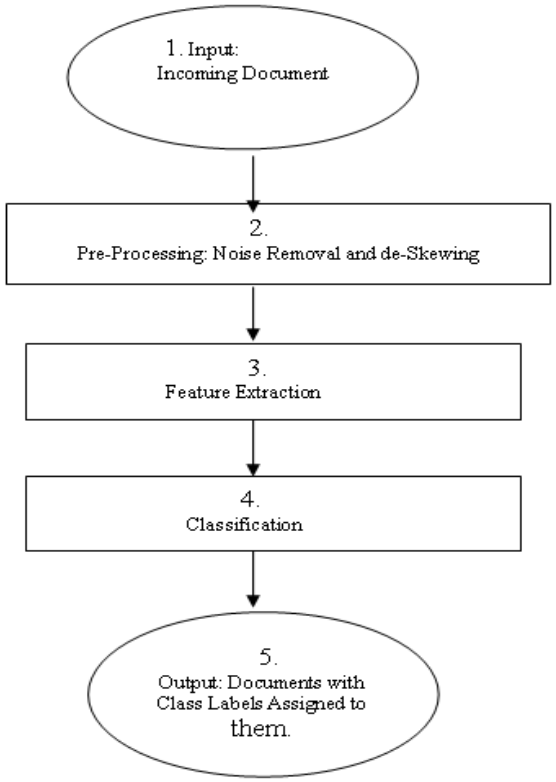
Text Processing
Deals with the textual components of a document image & its task are;

- Determining the skew (any tilt at which the document may have been scanned in the computer).
- Finding columns, paragraphs, textual lines, words, recognizing the text (Possibly its attributes such as size, font etc) by OCR.

Graphical processing
Deals with the non-textual elements (tables, lines, images, symbols, delimiters, company logo etc) Pictures are also included in this category; they are different from graphics in that they are often photographically or artistically generated.

3. DOCUMENT PROCESSING
Processing of document to extract their content in an automated fashion is essential task in all types of organizations for varied applications. Any document under processing is subjected to the following steps as depicted in figure 1.

1) The Pre-Processing Stage that enhances the quality of the input image & locate the data of interest.
2) The feature extraction stage that captures the distinctive characteristics of the document under processing.
3) The classification stage that identifies the document; groups the according to certain classes & helps in their efficient recognition.



"Figure 1"

3.1 Pre-Processing

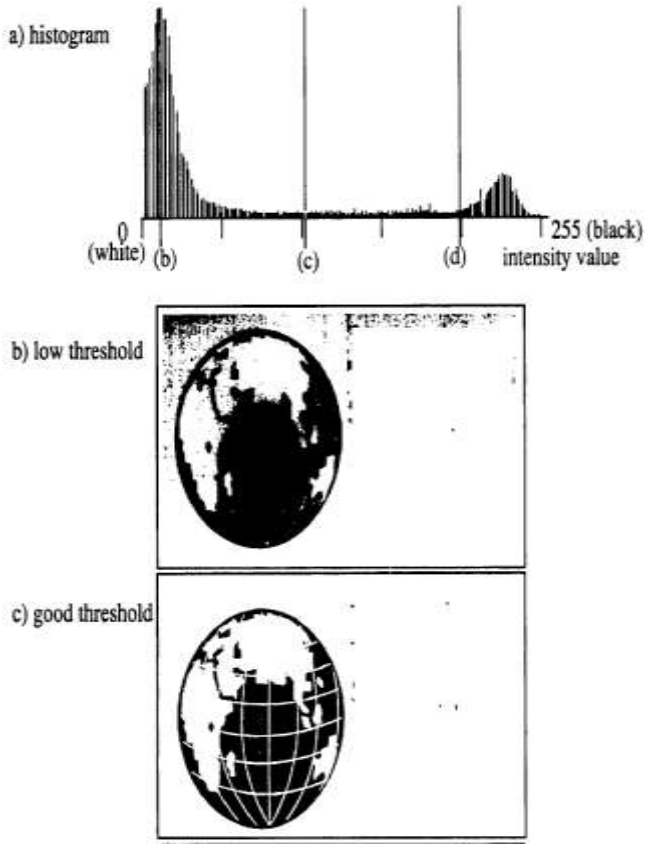
(Also known as Pixel-level processing or low-level processing) is done on the captured image to prepare it for further analysis. Such processing includes: Thresholding to reduce a grayscale or color image to a binary image, reduction of noise to reduce extraneous data, segmentation to separate various components in the image, and, finally, thinning or boundary detection to enable easier subsequent detection of pertinent features and objects of interest.

3.2 Image Acquisition

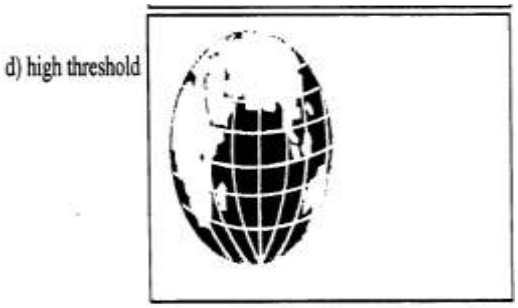
Acquire/obtain the image of document in color, gray level or binary format.

3.3 Binarization

Converts the acquired image to binary format, the objective of binarization is to automatically choose a threshold that separates the foreground and background information**.** Selection of a good threshold is often a trial and error process (see figure 2).A grey level of 128 is set as threshold. This becomes particularly difficult in cases where the contrast between text pixels and background is low (for example, text printed on a gray background).



"Figure2" Image binarization (a) Histogram of original gray-scale image. Horizontal axis shows markings for threshold values of images below. The lower peak is for the white background pixels, and the upper peak is for the black foreground pixels. Image binarized with: (b) too low a threshold value, (c) a good threshold value



"Figure2" (d) too high a threshold value

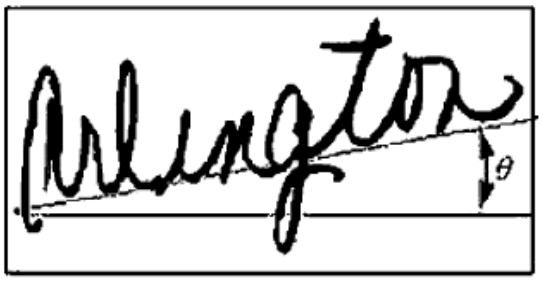
3.4 Noise reduction

The data extraction procedure often requires binarizing the images, which discard most of the noise & replace the pixel in the image, character & the pixel in the background with binary 0 & 1 respectively. After binarization, document images are usually filtered to reduce noise. For documents, more specific filters can be designed t take advantage of the known

characteristics of the text and graph components. A document to be scanned can itself be contaminated with dust or spots etc which constitute noise. Scanning itself can introduce some amount of noise. Noise is also due to the degeneration, ageing, photocopying or during data capture. In order to make it suitable for further processing, a scanned document image is to be freed from any existing noise. This can be achieved by a method known as image enhancement-this means improvement of the image being viewed by the machine or human. It includes contrast adjustment, noise suppression & many others. Smoothing operations in document images are used for blurring and for noise reduction. Blurring is used in preprocessing steps such as removal of small details from an image. In binary (black and white) document images, smoothing operations are used to reduce the noise or to straighten the edges of the characters, for example, to fill the small gaps or to remove the small bumps in the edges (contours) of the characters. Smoothing and noise removal can be done by filtering. Filtering is a neighborhood operation, in which the value of any given pixel in the output image is determined by applying some algorithm to the values of the pixels in the neighborhood of the corresponding input pixel. Various methods are applied to reduce noise. The most important reason to reduce noise is to obtain easy way of recognition of documents.

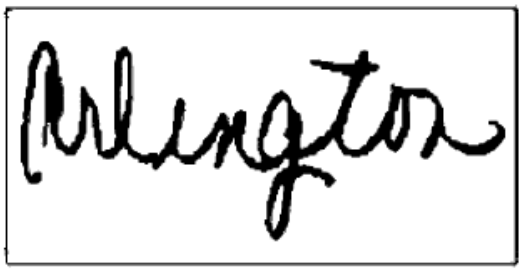
3.5 Skew Detection and Correction (De-Skewing)

Deviation of the baseline of the text from horizontal direction is called skew. This is result of improper paper feeding into the scanner. During the document scanning process, the whole document or a portion of it can be fed through the loose-leaf page scanner. Some pages may not be fed straight into the scanner, however, causing skewing of the bitmapped images of these pages. So, document skew often occurs during document scanning or copying. This effect visually appears as a slope of the text lines with respect to the *x*-axis, and it mainly concerns the orientation of the text lines; some examples of document skew are as follows in Figure 3;



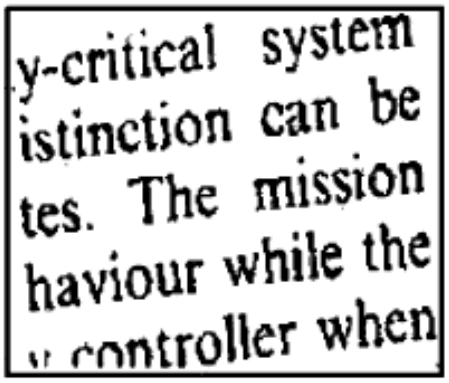
(a)

"Figure 3"(a) A skewed handwritten word



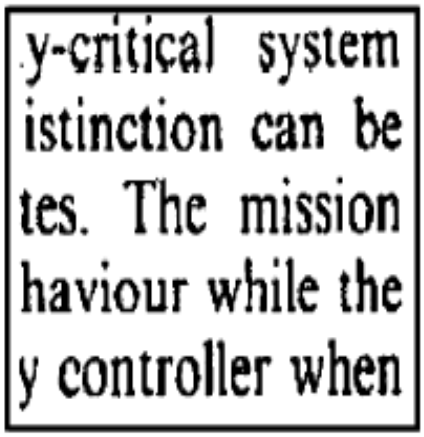
(b)

"Figure 3"(b) skew corrected handwritten word in (a)



(c)

"Figure 3"(c) skewed typewritten text



(d)

"Figure 3"(d) skew corrected image in (c)

Therefore skew detection is one of the primary tasks to be solved in document image analysis system, & it is necessary for analyzing a document before further processing. The document, if not placed properly on the scan surface, can introduce skew in

the resultant document image. Therefore, the document image may also need de-skewing.

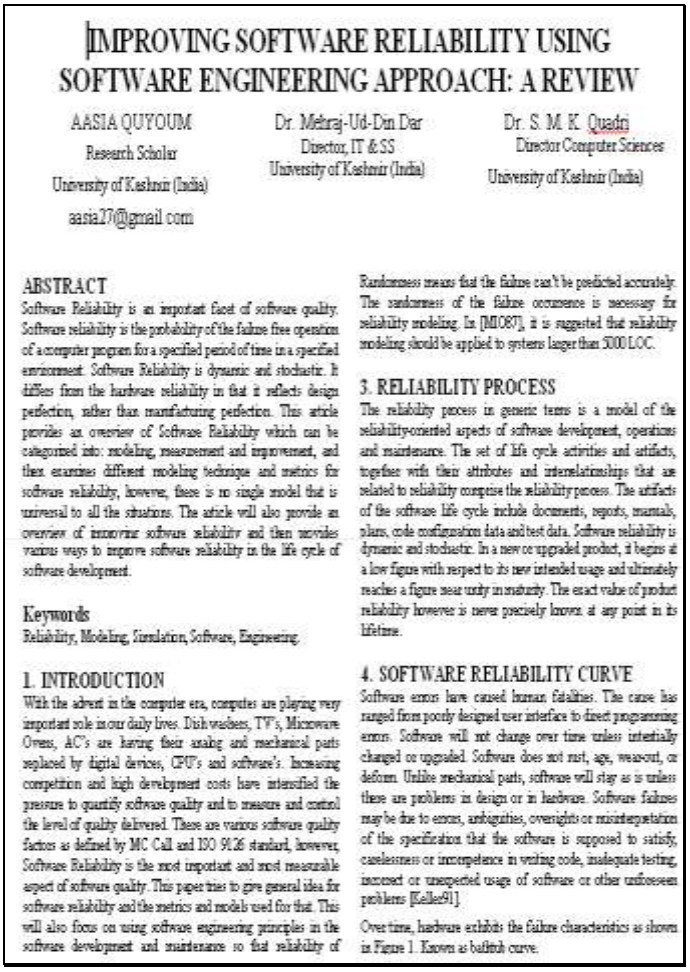
After skew detection, the image is usually rotated to zero skew angles, and then layout analysis is performed. Structural layout analysis (also called physical and geometric layout analysis

in the literature) is performed to obtain a physical segmentation of groups of document components. Depending on the document format, segmentation can be performed to isolate words, text lines, and structural blocks (groups of text lines such as separated paragraphs or table of contents entries). Functional layout analysis (also called syntactic and logical layout analysis in the literature) uses domain-dependent information consisting of layout rules of a particular page to perform labeling of the structural blocks giving some indication of the function of the block. (This functional labeling may also entail splitting or merging of structural blocks.) An example of the result of functional labeling for the first page of a technical article would indicate the title, author block, abstract, keywords, paragraphs of the text body, etc. Figure 4 a, b, c shows an example for the results of structural analysis and functional labeling on a document image [4].

Figure 4 shows the original document page is shown with results from structural and functional layout analysis. The structural layout results show blocks that are segmented on the basis of spacing in the original. The labeling in the functional layout results is made with the knowledge of the formatting rules of the particular journal.

3.6 Feature Extraction

Feature extraction involves the extracting the meaningful information from the document image. So that it reduces the storage required. The features that are extracted from whole image are known as the global features & the features that are extracted from blocks identified during segmentation or from subdivision (sub sectioning) of the document are known s local features. They can be divided into several categories: textural, geometric, component, structural and content based [15].The extractions of global and local features provide input to classification algorithm/techniques. One of the most important advantages of feature extraction is that; it significantly reduces the information (compared to original image) to represent an image for understanding the context of that image. Simple feature extraction methods, like calculating the difference between the minimum and maximum coordinates of the document image and the shape of the document obtained by comparison of breadth and length of the document image are of prime importance to acquire information regarding the document under process.



IMPROVING SOFTWARE RELIABILITY USING SOFTWARE ENGINEERING APPROACH: A REVIEW

AASIA QUYOUM
Research Scholar
University of Kashmir (India)
aasia27@gmail.com

Dr. Mehraj-Ud-Din Dar
Director, IT &SS
University of Kashmir (India)

Dr. S. M. K. Quadri
Director Computer Sciences
University of Kashmir (India)

ABSTRACT

Software Reliability is an important facet of software quality. Software reliability is the probability of the failure free operation of a computer program for a specified period of time in a specified environment. Software Reliability is dynamic and stochastic. It differs from the hardware reliability in that it reflects design perfection, rather than manufacturing perfection. This article provides an overview of Software Reliability which can be categorized into: modeling, measurement and improvement, and then examines different modeling technique and metrics for software reliability, however, there is no single model that is universal to all the situations. The article will also provide an overview of improving software reliability and then provides various ways to improve software reliability in the life cycle of software development.

Keywords

Reliability, Modeling, Simulation, Software, Engineering.

1. INTRODUCTION

With the advent in the computer era, computes are playing very important role in our daily lives. Dish washers, TV's, Microwave Ovens, AC's are having their analog and mechanical parts replaced by digital devices, CPU's and software's. Increasing competition and high development costs have intensified the pressure to quantify software quality and to measure and control the level of quality delivered. There are various software quality factors as defined by MC Call and ISO 9126 standard, however, Software Reliability is the most important and most measurable aspect of software quality. This paper tries to give general idea for software reliability and the metrics and models used for that. This will also focus on using software engineering principles in the software development and maintenance so that reliability of

Randomness means that the failure can't be predicted accurately. The randomness of the failure occurrence is necessary for reliability modeling. In [Mil087], it is suggested that reliability modeling should be applied to systems larger than 5000 LOC.

3. RELIABILITY PROCESS

The reliability process in generic terms is a model of the reliability-oriented aspects of software development, operations and maintenance. The set of life cycle activities and artifacts, together with their attributes and interrelationships that are related to reliability comprise the reliability process. The artifacts of the software life cycle include documents, reports, manuals, plans, code configuration data and test data. Software reliability is dynamic and stochastic. In a new or upgraded product, it begins at a low figure with respect to its new intended usage and ultimately reaches a figure near unity in maturity. The exact value of product reliability however is never precisely known at any point in its lifetime.

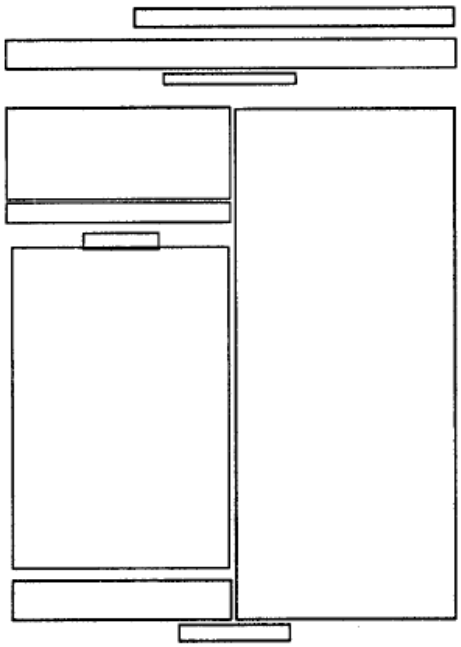
4. SOFTWARE RELIABILITY CURVE

Software errors have caused human fatalities. The cause has ranged from poorly designed user interface to direct programming errors. Software will not change over time unless intentially changed or upgraded. Software does not rust, age, wearout, or deform. Unlike mechanical parts, software will stay as is unless there are problems in design or in hardware. Software failures may be due to errors, ambiguities, oversights or misinterpretation of the specification that the software is supposed to satisfy, carelessness or incompetence in writing code, inadequate testing, incorrect or unexpected usage of software or other unforeseen problems [Keller91].

Over time, hardware exhibits the failure characteristics as shown in Figure 1. Known as bathtub curve.

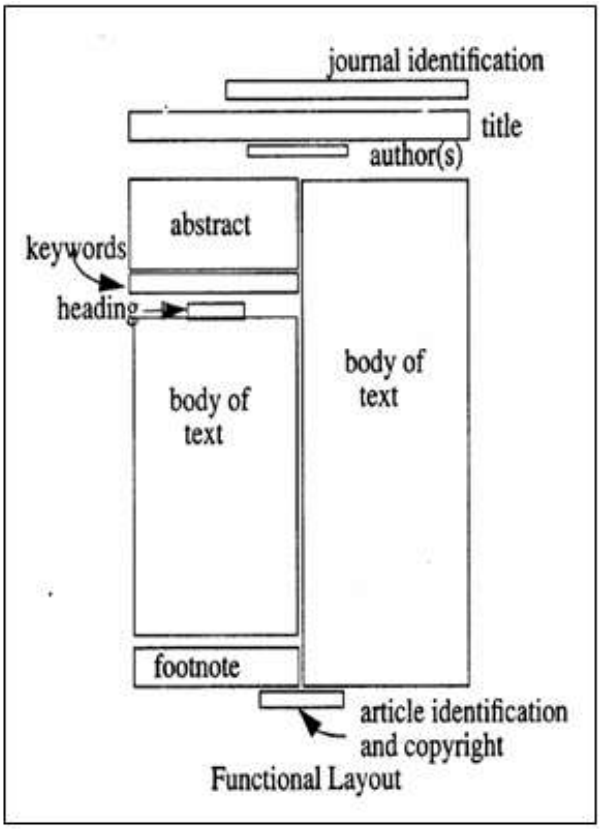
Original Document Page

"Figure 4 a"



Structural Layout

"Figure 4 b"



"Figure 4 c"

3.7 Classification

Image classification is a complex process and may be affected by many factors. The classification of document being processed is required for their efficient recognization as it reduces number of searches, easy recognization of document and also reduces the chance of error at different stages during processing. A classifier associates the document with class; labeling an observed document image according to the class, region in to which it falls. The classification stage identifies each input document image by considering the detected features like spatial arrangements with respect to one another, layout of document, size of the document, color of the paper, texture. The categorization (Indexing) of images greatly enhance the performance of document by filtering out the relevant document and the class to which it belongs. Classification of main document is done first followed by the sub-sections. A class prototype is stored in knowledge base .the incoming document is assigned to one of the classes, depending on the value of measure of nearness with the class prototype. This value is obtained by comparison of document under study and the class prototype. The document is assigned to the class with which highest value of the measurement is obtained.

Document classification use

Document classification is an important task in document Processing.

• Document classification allows the automatic distribution or archiving of documents. For example, after classification of business letters according to sender and message type (such as order, offer, or inquiry), the letters are sent to the appropriate departments for processing.

• Document classification improves indexing efficiency in Digital Library construction. For example, classification

of documents into table of contents page or title page can narrow the set of pages from which to extract specific meta-data, such as the title or table of contents of a book .

• Document classification plays an important role in document image retrieval. For example, consider a document image database containing a large heterogeneous Collection of document images. Users have many retrieval demands, such as retrieval of papers from one specific journal, or retrieval of document pages containing tables or graphics. Classification of documents based on visual similarity helps narrow the search and improves retrieval efficiency and accuracy.

• Document classification facilitates higher-level document analysis. Due to the complexity of document understanding, most high-level document analysis systems rely on domain-dependent knowledge to obtain high accuracy. Many available information extraction systems are specially designed for a specific type of document, such as forms processing or postal address processing, to achieve high speed and performance. To process a broad range of documents, it is necessary to classify the documents first, so that a suitable document analysis system for each Specific document type can be adopted.

4. CONCLUSION

The processing of documents for the purpose of discovering knowledge from them in an automated fashion is a challenging task and hence an open issue for the research community. In this article we provide brief summary of basic building blocks that comprise of document image processing system which modifies pictures to improve them (enhancement, restoration), extract information (analysis, recognition), and change their structure (composition, image editing). Today information technology has proved that there is a need to store, query, search and retrieve large amount of electronic information efficiently and accurately. So document image processing is very challenging field of research with the continuous growth of interest and increasing security requirements for the development of the modern society. Sequences of data preprocessing operations are normally applied to the images of the documents in order to put them in a suitable format ready for information extraction.

5. REFERENCES

[1] Casey, R. G., Wong, K.Y.,(July 1990) "Document Analysis Systems and Techniques, Image Analysis Applications", Image Analysis Applications, pp.1-35.

[2] Castleman, K. R., Digital Image Processing. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1979

[3] C.C.Chang and D.C. Lin.s, (1996) "A Spatial Data Representation: an Adaptive 2D-H string", Pattern Recognition Letters 17(1996) 175-185, Elsevier.

[4] Claude Faure, Nicorevincent, "Document Image Analysis for Active Reading", International Workshop on Semantically Aware Document Processing and Indexing, ISBN 978-1-59593-668-4, pp 7-14, 2007.

[5] Dr. Mehraj-Ud-Din Dar "Document image classification: A Cognition Based Approach", J&K Science Congress University of Kashmir, 25-27 July, 2006.

[6] Gonzalez, Rafael C. and Woods Richards E., (1999), "Digital Image Processing", Addison Wesley.

[7] Guru, D.S., (2001) "Classification of documents: An overview, the challenges and future avenues", NCDAR, Proceedings of the Pre-conference Workshop on Document Processing, 12th July, Mandya, India.pp28-34.

[8] H. Arai and K. Odaka. Form reading based on background region analysis. In *Proceedings of the 4th International Conference on Document Analysis and Recognition*. Ulm, Germany, 1997, pp. 164–169.

[9] K.Y.Wong, F.M Wahl, "Document Image Analysis System" IBM journal of research and development, pp 647-656, 1982.

[10] Nawei Chen · Dorothea Blostein, A survey of document image classification: problem statement, classifier architecture and performance evaluation, IJDAR (2007).

[11] O Gorman, L., Kasturi, R., (July 1992), "Document Image Analysis Systems", Computer, 25, pp.5-8.

[12] RANGACHAR KASTURI1, LAWRENCE, and O'GORMAN2, Document image analysis: A primer, *S¯adhan¯a* Vol. 27, Part 1, February 2002, pp. 3–22.

[13] Samet, H, (1990), "Applications of Spatial Data Structure", Addison-Wesley, Reading,

[14] Sonka Milan, Hlavac and Roger Boyle, (1999), "Image Processing Analysis and Machine Vision", Brooks/Cole Thomson Learning.

[15] T. Young, Gerbrands, "Fundamentals of Image Processing", Paper Back, ISBN 90-756, 9th January 2007.

[16] Ye-In Chang and Hsing-Yen Ann., (1999), "A Note on Adaptive 2D-H Strings", Pattern Recognition Letters 20(1990) 15-20, Elsevier.

[17] Y.Y.Tang and C.Y.Suen, "Document Structure: A Survey", in International Conference on Document Image Analysis and Recognition, pp 99-102, 1993.