Chapter 1

Introduction
Chapter 1

INTRODUCTION

1. MOTIVATION

In the recent past, Document Image Analysis (DIA) [1-5] for information extraction has gained immense importance. The variety of information conventionally stored on paper is being converted into electronic form for better storage and intelligent processing. This includes information extraction [1] from document images. In order to achieve this, documents are processed using image analysis [6-9] algorithms. Document Image Analysis is a subfield of digital image processing [6-9]. This involves conversion of document images to symbolic form for modification, storage, retrieval, reuse and transmission. Document image analysis facilitates the transition from bookshelves and filing cabinets to the paperless world. Document Analysis aims at transformation of any information presented on paper and adjust to human comprehension into an equivalent symbolic representation. This symbolic representation is accessible to any kind of computer information processing.

Part of this chapter is published in
The degree of perfection as compared with human being needs to be improved. Specialized tasks are involved in this process but technology has been able to solve this problem partially. The problem is solved to some extent with the help of high resolution scanning devices and increasing availability of computing power. This leads to accurate document analysis which facilitates high level of office automation, document analysis, document understanding [10,11], Optical Character Recognition (OCR) systems, text reading and so on.

Typically document image analysis differs from conventional image processing in view of format and information contents. The information in document images is more structured. The information is also presented in a natural language with the help of grammar and script. Thus document images are usually rich in formal information presentation. A document contains text blocks and images (graphics block). The text blocks may contain paragraphs of text in various fonts and sizes, titles or captions. It is very difficult to extract some information from graphics blocks as compared to that from text block. A text block can be converted to an editable text, if the respective script, font, character can be recognized. In the present information rich era, situations where the search results are needed to be at the finger tips are bound to occur. The two basic issues are to represent the bulky raw-data in the compact and interactive form and to retrieve relevant information from the database. According to the image content, images can be divided into two categories viz. Natural and Symbolic [1]. Portraits, fingerprints, aerial photographs, satellite images and x-ray, depict natural scenes or objects. Whereas postal addresses, printed articles, bureaucratic forms, engineering drawings and topographic maps represent symbolic objects. In [1] George Nagy has proposed the definition as,

"Document Image Analysis is the theory and practice of recovering the symbol structure of digital images scanned from paper or produced by computer."

The pictures of symbols are produced with high contrast, for easy reading. Hence most text and line art is essentially black-on-white. Color is applied if necessary, without fine tonal gradations. Photographs are reproduced as halftones. The linear signal analysis techniques based on frequency transforms are less prevalent in DIA than in computer vision and in natural picture processing.

Within the stream of raw input data meaningful objects and their relations must be detected. It is required that the document analysis task must be structured into several levels of interpretation and requires combination of bottom-up and top-down
approaches. Ambiguities are to be resolved only at higher levels. It needs to provide many alternatives with the goal for minimizing number of errors. This may lead the Document Analysis stream to a knowledge based system.

1.2 DOCUMENT ANALYSIS REVIEW

A document is a written or printed paper that bears the original, artificial or legal form of data and can be used to furnish decisive evidence or information required. We encounter a wide variety of documents in our daily life. For e.g. the documents used to communicate information in the form of letters and newspapers, archived documents for later validation or use. The conventional documents are now being replaced by electronic documents. The challenges associated with processing of such electronic documents are considerably different from that of information extraction from digitized or converted electronic documents.

1.2.1 Document categories

A document image may consist of textual and visual information, which are digitized versions of handwritten or printed paper documents. The broad categories of documents characterized on the basis of content or the generation mechanism are[12]

i) Hard and soft document

ii) Printed and hand-written documents

iii) Single language and multilingual documents

iv) On-line and off-line documents.

i) Hard and soft documents

A document in physical form having textual or graphical information is called a hard document. Soft documents are those which are created with the help of electronic devices. A large number of electronic devices allow the conversion of hard documents to soft documents. Most of these devices convert an analog signal (illumination or pressure) into a digital representation. Cameras, flat bed scanner, hand held scanner, stripe reader, keyboard are some of the popular digitizing devices. Soft documents include documents generated using markup languages, word processors, synthetic document images and scanned images. The soft documents can be either structured or unstructured. The information is presented in a well organized manner in structured documents. In case of unstructured document, the understanding with the help of computational models is major challenge. Problems in unstructured document processing involve image processing and pattern analysis tasks.
ii) Printed and Hand written documents

Printed documents are generated either mechanically or electronically from the existing data. These documents have the advantage that all the occurrences of any single character are apparently similar. The document image analysis systems aiming at recognition of such documents can make use of this favorable property. These documents are characterized by the deterministic nature of characters, their repetition patterns and well-defined layouts. Moreover such documents are of very high quality. Handwritten document processing and recognition is relatively difficult due to cursive scripts, variability of characters, separation of characters, skewness of documents, variability of strokes etc. Most of the handwritten documents understanding systems are developed for a specific narrow domain. Sometimes we may encounter hybrid documents where printed and hand written words are present together. These documents may also contain texts, tables, graphs and graphics.

iii) Single language and multilingual documents

Character recognizer is the important component of any document image analysis systems. A character recognizer converts a character image into an ASCII (American Standard Code for Information Interchange) or ISCII (Indian Standard Code for Information Interchange) text. The performance of recognizing algorithm deteriorates with the increase in number of classes and variability within a class. The accuracy achieved for one language cannot be extended to other language because of the alphabet size and variability. The complexity increases when multiple scripts are present in a document. India being country with diverse languages and scripts, the real life documents are multilingual. Many documents use English, the national language Hindi and a regional language for official and commercial purposes. For processing of such documents, identification of script and appropriate recognizing schemes are employed. Performance of character recognition system is improved if appropriate script is identified as a priori. It is very common to find handwritten or printed documents with isolated presence of foreign language words.

iv) On-line and off-line documents

The on-line documents find tremendous application with handheld devices and natural interfaces. The digitizer also provides the information along with the spatial and intensity content of the image. The conventional documents analysis focuses on the off-line documents. A character is considered as a single image in case of off-line documents and on-line documents represent characters as a sequence of strokes.
algorithms for processing of on-line documents utilize the information and map the problem into an ordered sequence analysis.

1.2.2 Application Areas

The field of document image analysis has many potential applications. Few of them are listed below.

i) Newspaper documents

ii) Form processing (Bank cheque processing)

iii) Postal mail processing (envelops and letters)

iv) Archival of existing documents

v) Legal document processing

vi) Speech applications (Reading aid for the blind - text to speech)

vii) Digital Libraries

i) Newspaper documents

Newspaper is the first mass communication media introduced by human being. Afterwards radio, television and internet have become popular mass communication media. In spite of all these, newspaper appears to be most popular and appealing source of information for many people. Newspapers are hard printed offline documents, with very rich information content. These documents are in existence since printing was invented. Currently a vast number of newspapers are available and it is very difficult to maintain them in a paper format. This attracts attention to the importance of digitization, understanding and archival of newspaper documents.

Typically a newspaper of size 24in x 36in scanned at 300 dpi with 8 bits per pixel requires around 2 mega bytes of storage space. Preserving the newspaper documents digitally, facilitates the preservation of endangered material (paper) and the creation of digital library services that allow full utilization of the archives. One should be able to search and retrieve all relevant information about a particular national figure with minimal effort from a large newspaper database. Document image analysis applications take care of such tasks by developing integrated systems that provide solutions to problems related to digitization and classification of newspaper images. Results of these systems can be used in information retrieval and content based indexing.

ii) Form processing

A form is a document which requests or collects information from a user in specific format. These are the most common documents which are to be handled in
organizations. For example, forms are used in handling deposits and withdrawal in banks, educational institutions, applications for jobs, railway reservations, competitive examinations, surveys and data collection. Every field in a form has its corresponding value. The extraction of the information related to each field in the form is a major concern in designing form understanding software. As forms are meant for very specific applications, one usually expects very high degree of accuracy and processing speeds. The problem of information processing from forms is more structured in nature. The recognition part is complex since the texts are usually handwritten. Form processing introduce challenges to address the issues like, leaving some of the fields blank and leaving spurious marks on the fields such as overwriting and erasures. Accuracy of the form processing depends on the design of forms. The better is the form design the better is the accuracy of form processing.

Forms contain many standard characteristics such as lines, boxes etc. The sequence of operations in the layout analysis depends on the prior knowledge about forms. Model driven and data driven approaches are adopted for form processing. If the form structure is unknown, a possible approach is that of classifying documents in a fixed number of classes and using the model corresponding to each document class to carry out the layout analysis.

iii) Postal mail processing

Regular mails are a major communication mode. A remarkable attention has been paid by the researchers to the sorting and classifying of the mails based on handwritten or printed addresses. An automated system can make this process accurate and fast. The concept of zip codes or PIN codes was introduced to increase the accuracy which reduces the delivery time of the mails. The recognition of handwritten or printed numerals is relatively simple as compared to the recognition of text. In India, the postal addresses are often multilingual and are written in cursive scripts without PIN codes. Due to this the problem has become extremely difficult to handle. Most of the times, locating the addresses and other details is very difficult. Many people also use their own abbreviations and spellings for the names of places. The postal address recognition system developed at CEDAR is a successful document image analysis system.[12]

iv) Archival of existing documents

This is an important application of document image analysis where an existing hard document is converted into a soft text form, for easy manipulation of the
document. The intelligent digitization of paper documents has tremendous applications in various domains. The judgment given by the courts fall under these types of documents. A proper digitization of the documents in the form of judgments can help the judiciary to provide speedy and better judgments. Initially the judgments were handwritten, later they were typewritten and in recent years the judgments are completely electronic. The recognition performance in judgment document can be improved by using domain knowledge because of the limited and special vocabulary used in official documents. But this will make the system domain specific and will increase the difficulty in developing a generalized system.

In India, document image analysis systems can be put to work for rapid digitization of abundant official and statistical information in administrative offices. These documents contain a lot of information in addition to text, in the form of tables, bar graphs, pie-charts. These graphical components of documents are different from images and need to be interpreted separately depending on the content and the information they want to convey. This gives rise to the problem of text and graphics separation from the document image and analysis of graphical sections of the image.

v) Legal document processing

The objective of legal text handling is actually to combine different types of texts in an effective way. Statutes, regulations, cases, precedents, legal literature, contracts are examples of documents that may have to be investigated together in order to solve a legal problem or even to be able to understand the practical meaning of a legal rule. This can be described as legal rule fragmentation: the necessary information is often scattered in different documents or even in different data banks, and the links among the required pieces of information are difficult to establish [13]. This problem is not particularly limited to the legal area only.

In the recent years a growing interest has been dedicated to the study of complex representation models, in which sophisticated search functions (typical of the Information Retrieval field) are integrated with navigation functions (typical of Hypertext systems) [13]. An effective combination of the benefits provided by the two models is expected to provide the best support tools for the localization of scattered information that is of interest for the user. The rapid expansion of the Internet, which has resulted in a rapidly growing worldwide hypertext, has provided additional momentum to the research in this area.
An economical problem which hampers the development of powerful models is the cost for the manual indexing/authoring task. It involves extraction from texts of auxiliary data suitable to encode various relevant aspects of text content (typically, cross-reference citations and concepts belonging to a pre-defined thesaurus or classification schema). This manual encoding activity is hard, error prone and time-consuming task. In legal field the importance of this problem can be easily understood by considering the extremely rapid growth of the overall document collection. Natural Language Processing (NLP) techniques are a key resource in order to overcome this economical problem.

vi) Speech Applications

Applications like document reading devices involve OCR followed by a text to speech (TTS) converter. The TTS converter converts the text obtained from OCR output of document image into sound signal [12]. These type of applications are very much helpful to blind and illiterate people. Further advancements in this area may introduce a language translation module between the OCR and TTS. This will help any person to understand a document in any unknown language.

vii) Digital Libraries (DLs)

The Document image analysis system aims at the transition from bookshelves and filing cabinets to the paperless world. The digital libraries[14-21] can provide useful data and metadata for research in automated document image analysis systems. The development of digital libraries is reaching technological limits due to the difficulty of automatic processing of digitized images of documents. The DIA open problems in digital library applications are identified at every stage of capture, early processing, recognition, analysis and retrieval. These are grouped as per the processing stages, where they are applicable in the construction of digital libraries. Critical discussion about the general and specific challenges for each class of document in DIA is covered in this chapter.

In the recent past information is captured electronically in digital form (e.g. ASCII, Unicode, XML) and the resulting data is made available on-line as digital libraries and other web resources. The information that is not yet converted to digital form is difficult to find, access, read, understand and otherwise reuse. Electronic “Digital Libraries” (DLs) [14-21] are designed to overcome these difficulties and provide scanned digital images of originally printed and handwritten materials. In fact,
Digital Libraries facilitate many ‘lost’ texts - out of print, deteriorated, mutilated, locked in archives, etc - into circulation.

The physical properties of good quality paper document include lightweight, thin, flexible, markable, unpowred (‘always-on’), stable, cheap, etc. The digital display devices such as desktop, laptop and handheld computers, plus e-Book readers, tablet PCs etc. used for digital libraries have many advantages. They are automatically and rapidly rewritable, interactive, connected (e.g. wirelessly) via networks to vast databases. Sellen and Harper [14,15,17] reported, “paper [remains] the medium of choice for reading, even when the most high-tech technologies are at hand.” For this they suggested the reasons:

1. Paper allows “flexible navigation through documents;”
2. Paper assists “cross-referencing” of several documents at one time;
3. Paper invites annotation; and
4. Paper allows the “interweaving of reading and writing.”

New technologies such as E-ink[17] and Gyron[17] facilitate electronic document display with more advantages of paper. The research is needed into user-interactions with displays during reading and browsing, in order to understand fully the challenges in the delivery of document images via DLs.

1.3 DIGITAL LIBRARIES ISSUES IN DIA

There are various reasons which justify an expensive digitization of documents from cultural and scientific heritage.[14,15]

**Long-term cost reduction of duplication and consultation:** Digitization is an expensive process that is becoming more competitive considering the cost of microfilm duplication and consultation. Microfilm viewers are very expensive and difficult to repair. Digitized documents can be reproduced indefinitely without loss of information, stored and exchanged using a large variety of supports and displayed on many types of devices. Digitization allows book-on-demand publishing systems to reprint old documents in high quality.

**Better access to rare collections:** Digitization projects primarily concern rare and old documents which can be accessed only by a limited number of students and researchers. The digitization of rare collections will improve their accessibility to a wider audience and will protect rare and fragile books from frequent handling. On-line digital libraries, freely accessible, will reduce the number of visitors and receptionists
in libraries which can focus their resources on their main objective: the preservation of documents.

**Digital Libraries provide new services:** Digital libraries have already developed new services for users like efficient information retrieval, document comparison, collaborative works, documents' edition tracking, etc. For digitized documents, these new services would not be possible without better automation in the retro-conversion process.

**Cultural heritage spreading and preservation:** On-line digital libraries are developing new cultural activities open to a wider public. This will promote the cultural heritage of a city, state or country all over the world. Digitization is a long-term investment because generated files can be handled, duplicated and exchanged electronically everywhere in the world. To prolong the usable life of electronic records, we must frequently change the support, the file format and upgrade the software and codecs.

Digital libraries facilitate services such as on-line consulting of rare documents, improved navigation capabilities, information retrieval and the ability to share knowledge with other readers. For this, it is needed to digitize all books of importance to human knowledge. Numerous research projects on digital libraries are dealing with the future tools of libraries for query, retrieval, analysis, management, accessibility, usage, archiving and preservation of information. A growing part of Digital Libraries are accessible in image mode because old manuscripts or ancient printed books cannot be processed automatically by an OCR system [1]. Most documents are still preserved in paper format which should be digitized and stored in a reusable electronic format. The digitization [1,14-17] process consists of capturing images of documents and the retro-conversion process defines all operations required to convert images into reusable data. Retro-conversion is achieved by indexing images, extracting metadata and converting into a file format suitable for the user needs. There are two different classes of problems.

**General problems:** The objectives of digitization, the functionality of the digital library according to the user needs, the finance allocated for the project and the limitations of available technology must be well defined. The research community should advice the technological choices and solutions for the unsolved problems. These choices concern all part of the digitization process starting with the image quality, color depth and resolution, the choice of digitizers, the image file format and the type of

*Document Analysis: A Text Based Approach*
compression (lossy or lossless), the image indexing system, the choice of database management system, the file format exchange for the publication on the web, the functionalities of the digital libraries and the ergonomics of the client interface.

Specific problems: The documents from cultural heritage have specific metadata. Their indexing requires the intervention of experts. The development of understanding systems to process automatically all type of digitized documents is difficult because each application needs very specific developments on demand. The solution consists of collaborating with libraries and users to define the information to be retrieved for both image indexing and/or image description. A difficulty for the development of digital libraries is the high cost of the digitization and retro-conversion processes. The cost of retro-conversion can be reduced by introducing efficient image analysis software starting from image processing basic tools to document understanding systems. Another difficulty is the reduced accessibility of digital libraries in image mode through the Internet. Digitized documents are represented by heterogeneous data (images / text / links / annotation / physical layout and logical structure) which is not fully managed by actual file formats. Users ask for a common file format suited for documents description and representation which can be queried by contents, edited, annotated, exchanged and progressively transmitted through a network.

In section 1.3.1 the general problems of digital libraries are discussed. Section 1.3.2 summarizes the specific problems for each class of documents. Section 1.3.3 deals with 'personal' and interactive digital libraries.

1.3.1 General problems of digital libraries

i) Digital Image quality

Image quality for book-on-demand publishing, web publication or image analysis is completely different and depends on the final objective[14]. It is noticed that a growing part of documents are digitized at insufficient quality for an automatic process. The important issue for the Document Image Analysis Community consists first to advice libraries on the quality of images required for an automatic processing. To save money and time, numerous projects digitize microfilms rather than the originals. The digitization of microfilms is an easy process which is faster and cheaper than using the originals.

ii) Image Capture

In order to build repositories containing part of cultural heritage, development of industrial processes, high quality software and low-cost hardware is required[14].
Book digitizers are the first step of a digitization line. They are highly sophisticated products which have several constraints such as the capture of high quality images, the physical preservation of the original documents during the image capture and the lowest possible cost of digitalization which depends on the speed of the capture and the ergonomic used.

iii) Analysis of Content

The analysis and recognition of the content of document images requires the full range of DIA R&D achievements: page layout analysis, text/non-text separation, printed/handwritten separation, text recognition, labeling of text blocks by function, automatic indexing and linking, table and graphics recognition, etc [14]. However, it should be noted that images found in DLs, tend to pose particularly severe challenges to today's DIA methods.

iv) Indexing, Retrieval & Summarization

The indexing and retrieval of document images are critical for the success of DLs. The JSTOR DL [15] includes over 12 million imaged pages from over 300 scholarly journals and allows searching on full text as well as on selected metadata (author, title, or abstract field). Most published methods for retrieval of document images first attempt recognition and transcription followed by indexing and search operating on the resulting encoded text.

1.3.2 Specific problems for each class of document

The diversity of image quality, complexity and contents is the main problem for the Document Image Analysis community (DIA) and the development of automatic digitization and retro-conversion lines [14,16]. The documents can be classified into four categories having almost the same contents and showing common difficulties for image analysis.

i) Medieval manuscripts, coins, seals...

These rare documents are difficult to digitize and must be handled with precaution. Their indexing needs great expertise that a computer vision system cannot achieve properly, but image analysis systems can retrieve useful information. The French Research Institute on Texts History (IRHT) is a CNRS laboratory which performs fundamental research on ancient manuscripts. The IRHT collects and digitizes rare manuscripts with the support of the French ministry of culture. Different metadata that image analysis systems can retrieve automatically are Illuminated objects. Main
ii) Early Printed documents of the Renaissance

The 16th century book interests all historians and book historians, because it is situated at a key moment in European history with the invention of printing by Gutenberg [16]. The contents of these books give interesting information on the social, religious and political rapid evolution. The contents of the books also changed during the 16th century, with the emergence of political, scientific, cultural, recreational and popular books which progressively replaced religious books. It was the beginning of the rationalization and codification of texts, typography for an improved reading comfort and a progressive standardization of spelling.

iii) Authors manuscripts from 18th to 19th century (Example of Montesquieu’s manuscripts)

Handwritten documents of 18th and 19th century authors constitute rare collections that are preserved in libraries or specialized institutes and are recently associated to innovating digitalization projects. Among those manuscript collections from the 18th and 19th centuries, an impressive and complete collection of French manuscripts is written by Montesquieu, a well-known author of the 18th century [16]. The Montesquieu handwritten collection has been marked by intensive uses and manipulations that is expressed by a poor document’s visual quality, with handwritten texts that are often erased with multi-writer annotations and corrections.

iv) Administrative documents of the 18th - 20th century (Archives, civil registration...)

Administrative archives, catalogs and civil registration documents interest many students, genealogists, sociologists, historians and all citizens. In many countries several millions of pages of administrative documents are difficult to access for the larger public. These documents interest every citizen who searches for information on the origin and the biography of ancestors. Archives are a collection of administrative acts, which concern the relations between citizens and the state. Every event (birth and death, civil marriage, divorce, military enrollment, tax payment, notaries acts...) have been recorded during several centuries. Archives are generally made of files and records properly classified. For image analysis, archives are represented by a large variety of documents, of different qualities, which show complex logical structures and contain many links and references, which must be retrieved for fine indexing [14].
1.3.3 Personal and interactive digital libraries

Research is recently underway in 'personal digital libraries,' with the aim of offering tools to individuals willing to try to scan their own documents and, handling imaged and encoded files, assemble and manage their own DLs. Thus there is a need for integrated tools for scanning, quality control and restoration, ensuring completeness, adding metadata, indexing, redisplay and annotation. As publicly available DLs gather large collections of document images, opportunities arise for collective improvement of the DL services[14]. For example, one user may volunteer to correct an erroneous OCR transcription; another may be willing to indicate correct reading order or add XML tags to indicate sections. In this way a multitude of users may cooperate to improve the usefulness of the DL collection without reliance on perfect DIA technology. Digitization and retro-conversion of libraries open new research fields for the community of Document Image Analysis. Document image understanding tools begin to provide solutions for the automatic retrieval of metadata, the document indexing by content and the automatic monitoring of the quality of image production during the digitization. Many libraries have already digitized a great quantity of documents without taking into account the quality of images required by image analysis. The DIA and DL communities must collaborate to develop efficient solutions for future digitization projects.

1.4 THE OPEN PROBLEMS IN DIA

Even today, there seems to be a technical infeasibility of automatically extracting highly accurate transcriptions of the content in a wide range of document images. The result is that most paper documents in a dangerously inconvenient state: compared with encoded data, they are relatively illegible, unsearchable, and unbrowseable. Thus a large fraction of our variety of paper documents seems to be threatened by a growing importance of 'purely digital' information. The DIA R&D community is working hard in order to face the challenges and to overcome these open problems in document image analysis.[22]

1.4.1 Capture

The capture [1,14,15,17] of document images often occurs in large-scale batch operations. The documents can be damaged or destroyed in the process, sometimes deliberately. It is thus urgent to design document scanning operations so that the resulting images will serve a wide variety of uses for many years. Image quality is most often quantified through the technical specifications of the scanning equipment, e.g.
depth/color, color gamut and calibration, lighting conditions, digitizing resolution, compression method and image file format. Such measures are vitally important means of quality control. Research is needed into goal-directed metrics of document image quality.

1.4.1.1 Scanner Specifications

Today Digitizing resolutions for textual documents typically range between 300 and 400 pixels per inch (ppi). 600 ppi is less common but is gaining acceptability as scanner speed and disk storage capacity increases. Researchers have suggested that grey-level processing will eventually allow OCR to read documents whose image quality is below average. Many documents are printed in color, and the costs of color scanning and of file storage and transmission are falling rapidly. Here are many research opportunities such as: does a particular scanning regime for modern books and printed documents – e.g. 300 ppi 24-bit color — reliably provide images (of text at least) which will support the best achievable recognition accuracy in the future as image processing methods improve? Or, should we argue for some more exacting scanning standard? It is clear that, for certain archival and scholarly uses [15,17], higher standards are already required. The Association for Information and Image Management (AIIM) [17] publishes standards for the storage of textual images. The DAS’02 Working Group on Digital Libraries [17] floated the idea of a tiered set of standards, to allow scholarly investigations sensitive to details of paper, ink, bleed through, evidence of age, marks of history of use, indicators of changing cultural context of the document, etc.

Table 1.1 - Scanning standards suggested by DAS’02 Working Group on DLs[17]

<table>
<thead>
<tr>
<th>Document Type</th>
<th>Scanning resolution in ppi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any document printed before the year</td>
<td>2000</td>
</tr>
<tr>
<td>1600</td>
<td></td>
</tr>
<tr>
<td>Document printed between years</td>
<td>600</td>
</tr>
<tr>
<td>1601-1800</td>
<td></td>
</tr>
<tr>
<td>Documents printed after 1801 and</td>
<td>400</td>
</tr>
<tr>
<td>later</td>
<td></td>
</tr>
</tbody>
</table>

1.4.1.2 Measurement & Monitoring of Quality

The technical specifications of scanning conditions should be preserved and attached (as metadata) to the resulting images. Exploratory research in this direction is
under way, but many questions are yet unanswered, e.g. How accurate will these estimates be? Can we estimate most of the image quality parameters that affect recognition? Will they run fast enough to be applied in real time as the images are scanned? [23]

1.4.2 INK-ON-PAPER

It is easy to enumerate the physical properties of ink-on-paper that assist human reading [15,17], e.g. lightweight, thin, flexible, markable, stable, cheap, etc. Even today paper remains the medium of choice for reading, even when the most high-tech technologies are available. The reasons are

1. Paper allows “flexible navigation through documents;”
2. Paper assists “cross-referencing” of several documents at one time;
3. Paper invites annotation; and
4. Paper allows the “interweaving of reading and writing.”

1.4.3 Initial Processing

A wide range of early-stage image processing tools are needed to support high-quality image capture. Image calibration and restoration must usually be specialized to the scanner. Image processing should ideally, occur quickly enough for the operator to check each page image visually for consistent quality. Tools are needed for orienting pages so text is right side up, de-skewing the page, removing some of the pepper noise and removing dark artifacts on or near the image edges. Has any text been unintentionally cropped? Are basic measures of image consistency — e.g. brightness, contrast, intensity histograms — stable from page to page, hour after hour? Are image properties consistent across the full page area for each image? These are the challenging problems in DIA which should be addressed. [23]

1.4.3.1 Restoration

The main purpose of document image restoration [1,15,17] is to assist fast & painless reading. For this, methods have been developed for contrast and sharpness enhancement, rectification (including skew and shear correction), super-resolution, and shape reconstruction. The DIA community has developed many algorithms — some very fast — for accurately detecting and correcting skew, shear, and other geometric deformations in document images.

1.4.4 Analysis of Contents

The analysis and recognition of the contents of document images requires the full range of DIA R&D achievements such as Page layout analysis [1], text/non-text
separation, printed/handwritten separation, text recognition, labeling of text blocks by function, automatic indexing and linking, table and graphics recognition, etc. Most of the DIA literature is devoted to these topics. A fruitful direction for DIA R&D is a search for tools that can reliably perform some tasks across the full range of naturally occurring documents. Such tasks might include:

1. Does an image contain any printed or handwritten text?
2. Does it contain a long passage (e.g. 50 words) of text?
3. Isolate all textual regions, separating them from non-textual regions and background;
4. Identify/segment handwritten from machine-printed text; and
5. Identify script (writing system) and language of regions of text.

1.4.4.1 Accurate Transcriptions of Text

The central task of DIA research has long been to extract a full and perfect transcription of the textual content of document images. The study reveals that no existing OCR technology, experimental or commercially available, can guarantee near-perfect accuracy across the full range of document images of interest to users.[23]

1.4.4.2 Determining Reading Order of Sections

Determining the reading order among blocks of text is of course a DIA capability critically important since it would allow more fully automatic navigation through images of text. This however remains an open problem in general, in that a significant residue of cases cannot be disambiguated through physical layout analysis alone, but seem to require linguistic or even semantic analysis. [23]

1.4.4.3 Tabular Textual Data

Detecting and analyzing tabular data is a problem which has received sustained attention by the DIA community [5,24,25]. It is of course harder in general than the analysis of images of body text; it appears however to be far easier than detecting and analyzing images of mathematical notation.

1.4.4.4 Representations

As there exists a variety of raster image file formats such as JPEG, PNG, GIF, TIFF, BMP etc. there are also many competing document file formats: ‘pure raster’ images, mixed image-and-text formats such as PDF, HTML, MSWord and PostScript, and ‘pure text’ formats such as ASCII/Unicode. The ability of a representation to express the result of every stage of document image analysis, from pure image all the way to completely encoded and tagged data is of interest to DIA researchers. Progress has been made toward this goal in academia and industry. [22]
1.4.4.5 Multivalent Documents

In multivalent documents [15,17] each page of each document is represented in multiple ways, often including: raster image, OCR-ed text, PDF, annotations, highlights, and hyperlinks. This requires fully automatic extraction of layout structure which is still a challenge to DIA R&D.

1.4.5 Presentation, Printing & Reflowing

The DIA community should help the design of these displays and should investigate versatile document image tiling algorithms. In many printed materials the author’s and editor’s choice of typeface, type size and layout are not merely aesthetic. Preserving all of these stylistic details through the DIA pipeline remains a difficult problem. One solution to this problem is, of course, multivalent representations [17] where the original image is always available as one of several views. Recently, DIA researchers have investigated systems for the automatic analysis of document images into image fragments that can be reconstructed or “reflowed” onto a display device of arbitrary size, depth and aspect ratio [17]. The document image reflowing systems work automatically on body text and still have some problems with reading order, hyphenation, etc. In most of these, some DIA problem needs to be solved [5,22].

1.4.6 Indexing, Retrieval & Summarization

The indexing and retrieval of document images are critical. Most published methods for retrieval of document images first attempt recognition and transcription followed by indexing and search operating on the resulting encoded text. An open problem is the effectiveness of OCR methods on very short passages, such as, in an extreme but practically important case, short fields containing key metadata (such as title, author, etc.). There has been, only a single DIA attack on the problem of summarization of documents by operating on images, not on OCR-ed text [17].

Non-textual content such as mathematical expressions, chemical diagrams, technical drawings, maps and other graphics have received sustained attention by DIA researchers, but it may be fair to say that search and retrieval for these contents is at a much less mature stage than for text.

1.5 HISTORY AND EVOLUTION OF DIA

Optical disks and tape cartridges were used for storing thousands of document images in the form of compressed images at reasonable price. Afterwards it was desirable to convert the document images to computer-searchable form. In 1999,
Document imaging, i.e., electronic document storage without sophisticated image manipulation became a billion-dollar business. Amazingly, DIA when considered as document image interpretation was a small part of it. The quality of production of documents through printing, copying and digitization, has enormous effect on the complexity of extracting information. The combined cost of conversion and processing determines where DIA can displace human data entry and correction.[1]

The digitized images can be stored in a single array format. Once the symbolic information is extracted and interpreted, the best format depends on the application at hand. Some of the formats are word processors (e.g., RTF), page composition software (LATEX, PDF) and web browsers (HTML). Some formats are specifically constructed for the manipulation of scanned documents (DAFS, XDOC and ODIL).[1]

Some of the early stages of processing scanned documents are independent of the type of document. Many noise filtering, binarization, edge extraction and segmentation methods can be applied to the document images.[1] Halftones require specialized treatment.

Once a document is segmented into its constituent components, more specific techniques are needed. Traditionally the DIA field has differentiated between mostly text and mostly graphic documents. Mostly text pages are separated into columns, paragraph-blocks, textlines, words and characters. OCR converts the individual word or character images into a character code like ASCII or Unicode. But text document should not be considered as merely string of symbols. Essential and additional information is conveyed by long-established conventions of layout and format, choice of type size and typeface, italics and boldface, and the two dimensional arrangement of tables and formulas. In order to capture the whole meaning of a document, DIA must extract and interpret all this subtle encoding. This is called metadata but it has different meanings in library science, web searches, programming and scripting languages. Specialized techniques are appropriate and affordable for high-volume text applications like envelopes, business letters, bank cheques and invoices.[1]

Engineering drawings, maps, music scores, schematic diagrams and organization charts are examples of mostly-graphic documents. The line drawings are decomposed into straight line and curve segments, junctions and crossings before higher-level components, such as dimensions, center-lines, and cross-hatching can be interpreted. Maps may require color separation and the association of text (labels) with map symbols. Line drawings contain lettering that must be located, isolated and
recognized. The information extraction from digitized photographs is an important research topic but since photographs seldom contain symbolic objects, our main concern in DIA is mistaking them for text or line drawings. At every stage of research, it is essential to be able to display the results of processing in a form suitable for human judgment. Accurate rendering of digitized pictures at various scales requires some care.

<table>
<thead>
<tr>
<th>Processing level</th>
<th>Mostly-text document</th>
<th>Mostly-graphics document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixels</td>
<td>Preprocessing, Representation, Noise reduction, Binarization, Skew detection, Zoning, Character segmentation, Script, language and font recognition, Character scaling</td>
<td>Preprocessing, Representation, Noise reduction, Binarization, Thinning, Vectorization</td>
</tr>
<tr>
<td>Primitives</td>
<td>Glyph recognition, Connected components, Strokes, Characters, punctuation words</td>
<td>Primitive recognition, Straight lines and curve segments, Junctions and nodes, Loops, Characters</td>
</tr>
<tr>
<td>Structures</td>
<td>Text recognition, Word segmentation, Text line reconstruction, Table analysis, Morphological context, Lexical context, Syntax semantics</td>
<td>Structure recognition, Text fields, Legends, Label attribution, Dimensions, Graphics symbols, Aerial and texture features</td>
</tr>
<tr>
<td>Documents</td>
<td>Page layout analysis, Text versus non-text, Physical component analysis, Logical components analysis, Functional components compression</td>
<td>Interpretation, Component recognition, Connectivity analysis, CAD/GIS layer separation, Database attribute extraction compression</td>
</tr>
<tr>
<td>Corpus</td>
<td>Information retrieval, Document classification and indexing, Search, Security, authentication, privacy</td>
<td>Database, CAD, GIS interface, Validation, Search, Update</td>
</tr>
</tbody>
</table>

Table 1.2 gives a typical document image analysis scheme[1]. The analysis is divided into five steps according to the granularity of the process such as pixels, primitives, structures, documents and corpus. The documents are grouped in two types mostly-text
and mostly-graphics. In case of mostly-text processing, top-down analysis attempts to find the larger components, like columns and paragraph blocks before proceeding to the text-line and word levels. Whereas bottom-up analysis forms words into textlines, lines into paragraphs and so on. As black pixel being the much smaller fraction of a line drawing, most techniques for graphics are bottom-up approaches.

1.6 CHALLENGES IN DOCUMENT IMAGE PROCESSING

A newspaper image requires several mega bytes of storage space to represent in binary form. The storage space can be drastically reduced if these document images are compressed. The representation of document images in compressed form allows easy manipulation of the content for many applications such as word processing and information retrieval. A document image analysis system can be employed to solve the problem of searching all relevant information about a particular individual from the huge newspaper archive. The major challenge in document image analysis is to locate text image blocks and tables, and defining appropriate algorithms for the same. Obviously this involves a number of preprocessing, segmentation and classification steps.

Beginning with the input image the document image analysis system initially tries to cancel out the noise effect and does the necessary geometric corrections. It is difficult for the computer system to identify the boundaries of an image and separate it from the textual region. Identification of table boundaries, graph boundaries and the text boundaries, and developing corresponding algorithms for table analysis, text, graphics and character recognition are some of the challenges which are to be faced in document analysis.

In the Indian context document analysis introduces many new challenges due to multilinguality and multiscrapping. Most of the systems developed are focused only on interpretation of a single script. Additionally, most of the Indian scripts have more characters in the alphabet. This increases the complication in pattern recognition. The presence of cursive scripts and matras also adds to the complication. Thus the document image analysis system developed in the Indian context should address the specific sections of the society. Handwritten documents from partially literate people may contain typographical and grammatical errors. The document’s paper quality and ink quality may also vary highly between users.
In the recent years, document image analysis has become a matured branch of research under digital image processing. The research on analysis and interpretation of images with Indian language content is being attended very recently. The International conference of document image Analysis and recognition (ICDAR) has witnessed a remarkable growth of work in Indian language documents. Other conferences like ICVGIP, KBCS, NCC and NCDAR also encourage research in the area of document image analysis.

1.6.1 Preprocessing

Preprocessing consists of a series of image-to-image transformations. It helps to extract the contents of the document.

i) Compressed representation

Run-length coding (RLC) and Freeman chain codes[1] were earlier used for compression. Later reduced terminal sequences of context-free grammars, coding on hexagonal meshes, production rules for sub blocks and filtered contours techniques for data compression were introduced.

ii) Filtering

The conversion of information on paper media to a digital media introduces noise. In image analysis noise is invariable. Many processes which involve transmission or change of media introduce noise.[1] The aim of noise reduction methods is to reduce the effect of noise for better performance of algorithms. Most popular methods for noise removal include mean filters and median filters where the pixel is replaced by mean/median in neighborhood. The image processing filters are designed either in spatial domain or in frequency domain based on the problem and computational complexity.

The group of pixels representing objects of interest is called foreground pixels and the remaining pixels are called background pixels. Finding the boundary between foreground and background is the process of segmentation. Image segmentation also partitions the image into more than two mutually exclusive and collectively exhaustive regions. The process of segmentation can be achieved in spatial domain or gray scale domain. Spatial segmentation is nothing but drawing a geometrical boundary between the objects in the image based on operations like edge detection, boundary identification etc. Where as grayscale thresholding divides the pixels into foreground and background based on a threshold gray value. The pixels on one side of threshold value are the foreground pixels and the other side of threshold is identified as
background pixels. This process is called thresholding. Depending on the way of working, the thresholding algorithms are broadly classified into global, local or adaptive techniques. [1]

iii) Binarization

For gray-scale images with information that is inherently binary such as text or graphics, binarization is usually performed first [2]. 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. This becomes particularly difficult in cases where the contrast between text pixels and background is low (for example, text printed on a gray background), when text strokes are very thin resulting in background bleeding into text pixels during digitization, or when the page is not uniformly illuminated during data capture. Many methods have been developed for addressing these problems including those that model the background and foreground pixels as samples drawn from statistical distributions and methods based on spatially varying (adaptive) thresholds. Whether global or adaptive thresholding methods are used for binarization, one can seldom expect perfect results.

In a noisy document image, the characters are separated from the background by performing filtering and binarization. The output quality at this stage is very critical for overall performance of the system [2].

iv) Skew detection and correction

Document images often get skewed at the time of creation. The skewed characters and documents can result in more misclassification and misinterpretation. Hence skew detection and correction forms a major preprocessing step in DIA, [1,2]. When a document is scanned, some amount of skew is introduced. It is necessary to ensure that the document is aligned properly. The skew can be associated with entire document or with individual characters. In either case, it is to be corrected before further processing. Almost all algorithms developed for skew detection are accurate on full pages of uniformed aligned printed text. The better algorithms are less affected by the presence of graphics, paragraphs with different skew, curvilinear distortion arising from photocopying books, large areas of dark pixels near the margin and few, short text lines. Projection profiles, Hough transform based techniques, nearest neighbor clustering are some of the popular methods used for skew detection and correction.

Projection profile is the histogram of the number of foreground pixel value in each scan line in the image. The projection profiles [1,2] provide clear evidence of
white lines as absence of black pixels in documents with horizontal text line and between lines white spaces. There are many fast and efficient algorithms developed using projection profiles which converge fast and accurately to the solution. One such algorithm divides the document into vertical and horizontal strips and takes projection profile for each strip. Then the skew is determined by measuring the average shift in zero crossings between strips.

Document image are preprocessed to provide black blobs in presence of text. In order to make the blobs into single lines they are thinned. These lines, when Hough transformed give rise to peaks in the parametric space. These peaks are identified and the skew of the document is estimated and corrected. Hough transform is computationally intensive. The Hough algorithm is seldom used for documents without extensive modifications.

In the nearest neighbor clustering method[1,2] to determine skew, neighbors of all connected components are found. The direction vectors for all the nearest neighbor pairs are plotted in a histogram and histogram peak is found to obtain the skew angle. The choice of the skew correction algorithm depends on the specific script and font.

1.6.2 Document Segmentation

In document analysis, the problem of separation of text and graphics and their identification and recognition is generally solved by making use of the differences in the properties of textual and image regions within the document. There are two views to look at a document, Top-down and bottom up.

The top-down view describes the document starting with a hypothesized format. The document is decomposed into textual and non-textual regions based on some assumptions. These segments are then broken up into finer sections recursively adapting the same technique. The method which assumes a document format, works efficiently and effectively for images of known document formats. In case of bottom up approach the document is built up from pixel level. The pixels in the document are grouped together with some constraint into smaller components. Then the components are grouped and the steps are repeated until the required area is totally covered. After that each of these segments are processed to identify the text, script and image content.

Smearing is used to group the areas of document image with similar features together. Smearing a document image horizontally results in converting the textual regions into black bands. The document images are smeared repeatedly until certain constraints are met. After smearing, the image results in a set of blocks. Features from
these blocks are then extracted and are classified into either text blocks or non-text blocks. To separate halftones from the images, finer classifiers are required. These methods are heuristic based and the results depend on the size of the font and other content of the image. The content in the segmented block is to be analyzed after segmentation. This analysis can be based on different features. The texture based feature is important one. The texture property of the block can be measured using filters like Gabor filter.

The structural information about the document image is obtained by horizontal and vertical projection profiles. This information is used to split the area into blocks and also can be used to identify what kind of data is present in the segmented blocks. The horizontal projection gives the horizontal boundaries of the image and text present. Similarly vertical bounds can be obtained by taking vertical projections. The intersections of the horizontal and vertical boundaries give the boundaries of blocks. This method directly yields the kind of data present in each block. Because of the variety of images further validation of the blocks is required. The method assumes the document to be in Manhattan layout. Projections can also be used in directions other than horizontal and vertical direction.

Agent based methods are used for document segmentation, by combining it with pyramidal image processing. The preprocessing for this kind of pyramid construction involves in constructing 8-connected-components and removing big objects which potentially belong to the category of images. The area histogram is used in this removal process. Thus the pyramid based algorithm builds the next layer with more of textual content rather than the image content. The components at each level are clustered based on some distance criterion. This method does not assume any kind of orientations in the textual regions of the image.

1.6.3 Document Decomposition

Some important classifications of components of a document based on the actions they perform are described as below.

i) Physical components

These are the components which make up the document in a macro way. In other words these components constitute the geometrical components in a document. Tables and images are examples of such components. For document classification and categorization problem, these are important features.
In physical component analysis one method analyzes the horizontal and vertical profiles of foreground pixels. Consider a document with text and gray-level or halftone images in it. Taking a horizontal profile of the foreground pixels gives a continuous strip of high foreground pixel count in the region where there is image and periodic count in the textual areas. Horizontal profile analysis gives the vertical extents of the contents of a document. Similarly by taking a vertical profile of the document, we get the horizontal extents of the documents. Combining these results gives the extent of geometrical components. After getting geometrical components separated, it is easy to observe what kind of region is it, text or image.

ii) Logical components

These are the primitive components which convey some semantics to the reader. Consider a postal application where envelopes with addresses written in different languages are to be separated into different categories for dispatch. Each address obeys a certain grammar. The extraction of logical components involves identifying an instance of the grammar on the document. This is achieved by decomposing a rectangular address blocks into sub images recursively by horizontal and vertical application of blocks grammars determining major logical components. Generally, the last line in address is city name followed by PIN code. The extracted PIN code can be sent to a digit recognition system which identifies the place where the letter is to be sent.

iii) Recognition of graphical components

In document image analysis recognition of graphics is an important area. It includes converting raster drawings to vector drawings, recognition of graphical primitives, recognition of shapes and symbols like logos, analysis and interpretation of diagrams like engineering drawings, logic diagrams, maps, charts, line drawings and tables. A vector drawing stores the image in vector representation rather than storing all the pixels of the image. For example, a line in the image is stored by its end points or its magnitude, direction and translation rather than storing all the pixels on the line. The conversion of machine drawings into vector drawings reduces the storage size to a great extent. The vectorized images are well suited for scaling and other transforms. This conversion makes the images editable and makes possible the content analysis.

Logo recognition is an important field of research which includes automatic grouping of official documents based on logos and other pictorial content that are present on the documents.
1.6.4 Multilingual documents

In multilingual document processing[2,12] the major problem is the determination of the location of the individual script and language content in the document image. After separating text regions into blocks from the image based on known scripts, text blocks can be further processed and characters can be directly sent to the corresponding OCR module for further processing.

Many general strategies are followed in separating and recognizing text segments in a document image. These are based on either the spatial domain or frequency domain characteristics of the scripts. Linearity, curvatures form the basic characteristics in spatial domain. In frequency domain, the response to various band pass filters are employed for the same.[26,27]

Optical density in text images provide clue for script recognition. This feature was applied in [28]. Another language set specific method is based on the projection profile analysis of the individual components. [29] uses statistical, topological and stroke based features in addition to the projection profiles. This method requires reliable segmentation of the characters. [30] uses script specific templates created for different languages in context and the input image is segmented using these templates. A direction distance based classifier initially takes the connected component and analyzes their neighborhood for recognition of scripts. Considering each pixel in the connected component as origin, a few imaginary lines in predecided directions are taken, and a number of black pixels along the lines at specific distances are chosen. This is called as direction distance histogram. The addition of these histograms for all the components yields the feature set of that specific region. This is matched to the existing database using Mahalanobis or Euclidean distance.

The features such as number of upward concavities, number of downward concavities, crossing counts are used for script identification. Some algorithms consider each language region in the document as a texture and apply texture segmentation procedures. One such reliable texture segmentation procedure is based on Gabor filters. A Gabor filter is a band pass filter characterized by the centre frequencies and bandwidths in the transformed domain. Here input image is passed through a set of Gabor filters and the outputs are used for computation of features.

1.7 Text Document Analysis

There are two main types of analysis that are applied to text in documents.[2] One is optical character recognition (OCR) to derive the meaning of the characters and
words from their bit-mapped images, and the other is page-layout analysis to determine the formatting of the text, and from that to derive meaning associated with the positional and functional blocks (titles, subtitles, bodies of text, footnotes etc.) in which the text is located. Depending on the arrangement of these text blocks, a page of text may be a title page of a paper, a table of contents of a journal, a business form, or the face of a mail piece. OCR and page layout analysis may be performed separately, or the results from one analysis may be used to aid or correct the other. OCR methods are usually distinguished as being applicable for either machine-printed or handwritten character recognition. Layout analysis techniques are applied to formatted, machine-printed pages. A type of layout analysis, forms recognition, is applied to machine-printed or handwritten text occurring within delineated blocks on a printed form. In some cases it is necessary to correct the skew of the document which is typically a result of improper paper feeding into the scanner. Skew estimation and layout analysis are discussed briefly in this section. General approaches to OCR are presented in the next section.

1.7.1 Skew Estimation

A text line is a group of characters, symbols and words that are adjacent, relatively close to each other, and through which a straight line can be drawn (usually with horizontal or vertical orientation). The dominant orientation of the text lines in a document page determines the skew angle of that page.[31-38] A document originally has zero skew, where horizontally or vertically printed text lines are parallel to the respective edges of the paper, however when a page is manually scanned or photocopied, non-zero skew may be introduced. Since such analysis steps as OCR and page layout analysis most often depend on an input page with zero skew, it is important to perform skew estimation and correction before these steps. Also, since a reader expects a page displayed on a computer screen to be upright in normal reading orientation, skew correction is normally done before displaying scanned pages. A popular method for skew detection employs the projection profile. A projection profile is a histogram of the number of ON pixel values accumulated along parallel sample lines taken through the document. The profile may be at any angle, but often it is taken horizontally along rows or vertically along columns, and these are called the horizontal and vertical projection profiles respectively. For a document whose text lines span horizontally, the horizontal projection profile has peaks whose widths are equal to the character height and valleys whose widths are equal to the between-line spacing. For
multi-column documents, the vertical projection profile has a plateau for each column, separated by valleys for the between-column and margin spacing. The most straightforward use of the projection profile for skew detection is to compute it at a number of angles close to the expected orientation [39]. For each angle, a measure is made of the variation in the heights along the profile, and the one with the maximum variation gives the skew angle. At the correct skew angle, since scan lines are aligned to text lines, the projection profile has maximum height peaks for text and valleys for between-line spacing. Modifications and improvements can be made to this general technique to iterate more quickly to the correct skew angle and to determine it more accurately [2]. Several other methods for skew correction have also been proposed [40-43].

1.7.2 Layout analysis

The image is usually rotated to zero skew angle after skew detection and then layout analysis is performed. Structural layout analysis [2] (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 (Dengel et al 1992) (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. See figure 1.1 for an example of the results of structural analysis and functional labeling on a document image. [2]
Fig. 1.1 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 (IEEE Trans. Pattern Anal. Machine Intell.). (Courtesy [2])

Structural layout analysis can be performed in top-down or bottom-up fashion [44]. When it is done top-down, a page is segmented from large components to smaller sub-components, for example, the page may be split into one or more column blocks of text, then each column split into paragraph blocks, then each paragraph split into text lines, etc. For the bottom-up approach, connected components are merged into characters, then words, then text lines, etc. Alternately, top-down and bottom-up analyses may be combined.

1.7.3 Hierarchical Structure of A Document

The regions in a document can take the form of text, graphics, or half-tones, and can be of any shape or size. Document decomposition provides specialized processing for such individual components. The page decomposition can be divided according to the nature of extracted structure. In order to facilitate OCR the text is to be separated from non-text. Further processing involves partitioning text into columns and paragraphs, non-text into line art and half-tones and demarcating tabular regions. The
correct descriptions of the region may not always be obtained at the pixel or component levels. In case of table interpretation, a valid decomposition may label a table region appropriately, but depending on the complexity of the model, a structural analysis may require a more complete description of the column, spacing, and separator components. Hence we can say that the decomposition is not complete, but that it divides the document into components which act as a guide to the interpretation process.[45,46]

For general document understanding problems, in which little is known a priori about the contents of the document, the process of decomposition, derivation of document class, and logical component labeling are interdependent. Beginning with a candidate decomposition of the document, it is possible to establish a hierarchy of abstraction which extends from the physical entities (syntactic components) up through the logical entities (semantic labels). The analysis task can be viewed as the derivation of a meaningful instantiation of this hierarchy based on information about the layout of the document and a model space which describes valid structured and logical document organizations.[45,46]

The structural analysis of documents involves more specifically the derivation of the logical or semantic meaning of a set of salient fields or regions within a document. In general the problem involves attributes and structural relationships of the document to label document components within the contextual rules dictated by the document class or type such as memo, letter, journal article, newspaper, etc. The analysis of structured documents relies on three components:

i) A meaningful decomposition of the document into primitive physical entities,

ii) The association of the document with a class of documents which can be used to guide the analysis, and

iii) The labeling of individual components as to their logical meanings and their relationships with each other.

The hierarchical structure of a document can be viewed in a logical and geometric way.

i) Logical Structure

A document is normally composed of several articles, each of which consists of a title, an abstract, subtitles and paragraphs. They are connected to each other logically in a hierarchical structure. For example, the title dominates the abstract, chapters and
sections, while subtitles dominate paragraphs. Thus, a document has a logical hierarchy.

ii) Geometric Structure

A document image is composed of several blocks, each of which represents a coherent component of the document. One coherent component corresponds to a set of text lines with the same typeface and consistent line spacing. The geometric structure means the geometric relationships between blocks.

A document image is broken down into several blocks and a geometric structure tree whose nodes represent a set of blocks is constructed. This approach first extracts words from a document image, which are then merged into text lines. Text lines are then combined into blocks which usually correspond to paragraphs. A geometric structure is generated according to the parent-daughter relationships between blocks. These relationships are established by examining the column containing a block and its vertical position. [45]

The run-length representation is more efficient than a bit-map representation, especially in image processing by software. Accessing pixels in a bit-map image is usually very time-consuming because a general-purpose computer is not designed to deal with bit-map images stored in a byte representation. The run-length representation for document image processing yields high speed processing. [45]

In the text line extraction procedure we define a segment as a rectangular area which circumscribes a text line or a part of it. The text line extraction process is divided into four sub processes. The first is to extract adjacent connected components as a segment. The second classifies the segments into text lines, figures, graphics, and so on. The third is a merging process for adjacent segments which are classified into text lines. The last is another merging process for the segments in the same column defined by the column boundaries. Words are usually extracted by both the first and second sub processes, and text lines are obtained in the third sub process. The fourth process is added to cope with cases where a long blank between words prevents the words from being merged into a text line[45].

1.7.4 Text Categorization

Text categorization deals with the automated assignment of documents into predefined categories or message types. We usually classify business letters into given message types such as order, offer, enclosure, advertisement, etc. The problem is that
OCR text can be noisy. Thus, the categorization techniques should be robust towards different kinds of recognition errors. In general, categorization of documents allows the automatic distribution or archiving of letters and is also an excellent starting point for higher-level document Analysis. [45]

1.7.5 Image Recognition

The image recognition phase involves recognition of the pixel image in each segment. If the segment contains a properly segmented character image, then, ideally, the output is the set of character labels and relative confidence values that a human would assign if asked to identify the image in isolation. As with segmentation, image recognition does not always live up to its ideal. An “8” which a human would have no trouble recognizing in isolation might be deemed an ambiguous “8” or “B”. In worst case it may be recognized as an unambiguous “B”. A split “d” might be recognized as “cl”. An “m” join might be recognized as “m”. The recognition module is to be tuned to make the best decision from too many choices for contextual clues during the ambiguity resolution phase. Image recognition is divided into three phases: feature extraction, classification and re-segmentation.

Feature extraction describes the image in each segment as a vector of fixed length. Classification outputs the character label it believes is represented by the feature vector. If it is unsure it outputs a set of choices and associated confidences. If it has reason to believe the segment might contain a join or split, it passes that opinion on to the re-segmenter. Classification is trained off line. Re-segmentation takes care of suspected joins, suspected splits, and re-circulates the new segments back to feature extraction. [45]

1.7.6 Character Segmentation

Character segmentation decomposes a sequence of characters into individual symbols. Text lines are obtained in the document analysis and document understanding phase. The character segmentation phase extracts characters from these text lines. This procedure consists in extracting and recognizing characters. After reviewing available literature, we conclude that there are three “pure” strategies for segmentation. [45] The elementary strategies are:

i) The classical approach, in which segments are identified based on "character-like" properties. This process of cutting up the image into meaningful components is given a special name, Decomposition or
Dissection. In this the input image is partitioned into sub-images having “character-like” properties.

ii) Recognition-based segmentation, in which the system searches the image for components that match classes in its alphabet. These methods rely on the integration of segmentation and recognition.

iii) Holistic methods, in which the system seeks to recognize words as a whole, thus, avoiding the need to segment into characters. These methods avoid segmentation by recognizing entire words as units.

In strategy i), the criterion for good segmentation is the agreement of general properties of the segments obtained with those expected for valid characters. Examples of such properties are height, width, separation from neighboring components, disposition along a baseline, etc. In strategy ii), the criterion is recognition confidence, perhaps including syntactic or semantic correctness of the overall result. The strategy iii), Holistic revert to the classical approach with words as the alphabet to be read.

1.7.7 Single Character Recognition

In document analysis, main task is the reconstruction - classification - of the textual information.[45] Classification is an act of abstraction. It establishes a mapping from a normally high-dimensional feature space into the discrete space of class labels. This mapping in case of character recognition is the inverse of printing or writing. The mapping is many-to-one and should be capable of producing the correct label irrespective of what the current character looks like. The dominant problem is to cope with the variability in character image appearance. The regions in the feature space in which the feature vectors of one character class appear are complex shaped regions which must be mapped into one and the same point in the space of character labels.

Variations in appearance may be caused by a number of different sources. Some may be modeled by regular transformations. It is a generally accepted fact that character class membership is independent of character image size - at least within certain limits. There are some other influences which leave the character meaning unchanged - such as translation, rotation, slant, and stroke width. Even if there are exceptions requiring additional measures, it sounds reasonable to look for features which are size-invariant, slant-invariant, etc. The simplest way of providing features with such invariance properties is to introduce normalization procedures between the step of gaining the first set of raw features and the recognition step.
The total classification process is divided into preprocessing and recognition, and implements the mapping from the raw image representation to character label in two sequential mapping steps. The positioning of the border between preprocessing and recognition is an engineering design decision. Also, extreme cases are conceivable in which each one almost totally replaces the other. In a reasonable design preprocessing applies normalizing transformations and reduces certain well-defined variations as far as possible.

1.7.8 Desired Features

In order to meet the market requirements, additional capabilities are required to build into the OCR system. Document analysis is the one primary component of the system. Its purpose is to capture the structure of a document and incorporate that structure into the OCR output so that, when viewed as an ASCII file or in a target word processor or spreadsheet, its formatting reflects the formatting of the original document.[45] This involves

i) Labeling the components of the document, such as paragraphs, tables, lists, headers, and footers;

ii) Establishing the reading order of the components;

iii) Analyzing the geometry of the page, for example, determining whether a title is centered or establishing the width of a column;

iv) Partitioning the text into regions that share the same font attributes, such as regions of italic, bold and underlined text and regions that share the same point size.

The appropriate formatting codes (e.g. centering codes, hard and soft new lines) are then passed to the application software.

1.8 Significance of the thesis

Document analysis helps in transforming any information presented on paper and addressed to human comprehension into an equivalent symbolic representation accessible to any kind of computer information processing. The degree of perfection compared with human literacy is an open goal. Actually, only a certain number of rather specialized tasks belonging to the above discussed category of problems are seriously approached and can be considered as partially solved. The document image analysis has captured remarkable position only in certain market segments. From our point of view the most dominant is postal automation, followed by form reading for
banking and administration purposes and page reading for general text input. However, with the advent of inexpensive high-resolution scanning devices and the rapidly increasing availability of computing power at the office workstation, fundamental prerequisites for document analysis are fulfilled. We are convinced that in the coming years document analysis will play an increasing role in the area of office automation.

Analysis of document images for information extraction has gained immense importance 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 algorithms. Locating text image blocks and tables, and defining appropriate algorithm is the major challenge in document image analysis [1,2].

The range and volume of publications such as newspapers, magazines, and various types of manuals are continuously increasing. At the same time, various computer-aided text processing techniques, such as desktop publishing, text data base management, and machine translation, have become possible. Automation is widely demanded in the keyboard input area, which was conventionally manual, where large amounts of documentation must be converted into a computer-readable form for data entry. A text reader meets this need. A text reader automatically analyzes each page of a document and recognizes characters on the page for input to a computer. It is important for such a text reader system to have the ability to deal with various kinds of document layouts and omni-font characters.

Many research topics are predominating in the analysis of documents. The digitization [2,22] (scanning or capture using other camera-based devices) of such documents must deal with requirements for detail preservation of both foreground and background information. The choice and availability of image acquisition equipment directly affects the digitization parameters and the overall image quality. In many applications the safe handling of fragile documents is of primary concern. Another obstacle is the presence of a multitude of different types of artifacts mostly due to degradations of the ink-based content (e.g., ink seepage, show-through etc.). The digital enhancement [3] of the captured documents is a research topic that has attracted considerable attention but at the same time remains a significant open problem.

The next category of challenges relates to the analysis of the layout and that of individual types of regions of interest [1,2]. The extraction of subsequently added
entities such as annotations and stamps as well as the detection of decorative regions are examples of regions of interest. The analysis of color (if present) is particularly important for both the differentiation of such entities from the main textual content and their accurate preservation. The recognition [1,2] of the content of different entities on the document is a major area of research and a particularly challenging one for historical documents[12]. Recognition objectives can vary from obtaining some form of encoding (useful for database retrieval at least) to full extraction of concepts based on linguistic analysis of text and domain-specific knowledge. The level of recognition that is actually possible is determined (in decreasing order) by whether the document is printed, typewritten or handwritten. At this stage the recognition of graphical components in terms of their decomposition into primitives is also of interest. Another very important consideration for documents is related to their use once they are digitized [1,2]. Depending on the type of document and the level of information that has been possible to extract from it, there are various possibilities for its representation and retrieval mechanisms [3]. There are several additional issues regarding post processing. The technique which is based on the comparison of shape-related features between the image of the query keyword (extracted from the document itself) and the individual images of the remaining words in the document is referred as word spotting [1,2]. It is very often necessary to capture knowledge and semantic information from experts such as non-technical experts (e.g., archivists, curators and historians) as well as provide them with user-friendly means of specifying, interacting with, monitoring and validating the analysis process and its results.

The task of Document Analysis usually encounters various types of input sources such as Printed character, Text, Drawing images, Mapped images, Magazine objects, Newspaper objects, Color objects, Photographs etc. The aim is to improve the range and volume of publications such as newspaper, magazines, various types of manuals, other documents which needs to be processed through various computer aided text processing techniques. Large amount of documentation needs to be converted into a computer readable format to avoid data entry. Document Analysis satisfies this need. The methodology followed is in following sequence.

1) The Document Analysis
2) Document Understanding
3) Character Segmentation/recognition
The research work presented here is for first stage that is document analysis. It is a component which decomposes a document image into several consistent items which represent coherent components of the documents such as text lines, photographs and graphics, without any knowledge of specific format. The Document Understanding [5,10,11] is a component which extracts the logical relationship between these items. Character segmentation/ recognition [47,48] is the component which then extracts characters from text line and recognize them. The document analysis and document understanding components need to be robust enough to cope with multicolumn and multi-article documents including graphics and photographs. The character segmentation/recognition component needs to have the ability to read omni font characters which might touch each other. As the application of existing technology to the analysis of documents depicts weaknesses, it is required to develop novel and more robust methods to cope with this challenging problem. A variety of serious technical obstacles prevent imaged paper documents from playing the useful roles in DIA.

The research is carried out here in document analysis component to make it robust enough to include various items like multi-column, multi-article, multi-font, multi-orientation and multi-lingual.

1.9 Outline of Thesis

Chapter 2 is dedicated to the literature survey. A detailed review of the various skew estimation techniques is given in the first section. Document structure analysis algorithms are reviewed the subsequent sections. A complete survey of the physical layout analysis and logical layout analysis techniques is given.

Chapter 3 explains the methodologies for document image analysis reported in the literature. Text document analysis and page segmentation problem is discussed. A complete overview of various page segmentation algorithms is also given in this chapter.

In chapter 4 an analysis of morphological document image segmentation is given. This chapter introduces the mathematical morphological operators and their properties. The role of structuring element in various morphological operations, basic shapes and types of structuring elements are also covered.

Chapter 5 deals with the design and implementation of the proposed morphological document image segmentation technique. The design parameters and design aspects are discussed. The proposed model for morphological document image
segmentation is presented. Lastly, implementation of the proposed technique by using various types of document images from NIST database (National Institute of Standards and Technology database) and other document images is presented in this chapter.

Chapter 6 presents performance evaluation, results and discussion. Results of all experiments performed in this research work along with statistical data are presented. The results obtained are also expressed in graphical form. The performance metrics which are used to evaluate quality of resultant document images in comparison with original document images are also analyzed and discussed.

In chapter 7, the conclusion and contribution of this research work is given. The future scope for the presented research work is also discussed.
References:


22. M.P. Dhole, Dr. V.M. Thakare, Dr. K.V. Kale, “Some Open Problems In Text Document Image Analysis”, Proceedings of National Seminar On Information
Technology Enabled Services (ITESS - 2009) organized by Department of Computer Science and Information Technology, Shri Shivaji College of Arts, Commerce and Science, Akola. 11th April 2009.


