International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

A Survey on Document Image Analysis

Dr. S. Vijayarani[1], Ms. A. Sakila[2], Ms. M. Deep[3]
[1]*Assistant Professor Department of Computer Science Bharathiar University Coimbatore*
[2]*Research Scholar Department of Computer Science Bharathiar University Coimbatore*
[3]*PG Student Department of Computer Science Bharathiar University Coimbatore*

Abstract*: Nowadays document images are becoming very popular and it is using in digitalized libraries and organization. Paper documents can be converted into digital form by using digitization equipment's. Digitalized document images are more compact to store in the computer memory and it becomes inexpensive. These document images also sent and receive via email, fax and other web usages and it simply view and print by users. Information in these document images are more structured and presented in a natural language with the help of a grammar and a script. This paper discusses document image analysis, applications of document images, challenges and issues for handling document images.

Keywords: Document Image analysis, Document Image Retrieval, Applications of Document* Image Analysis, Issues in *document images

I. INTRODUCTION

In document image processing, the paper documents are initially scanned and stored in the hard disk or any other required location, this scanned document named as a digitalized documents [1]. It is very popular in digital libraries and digitalized organization. Generally these document images also contains text and symbols and objects, for example book pages, postal addresses on letters, engineering drawings, sheet music, map and various types of document images. In the document image results will be in electronic format, through which the document will makes simple and easy to access. It comprises a set of simple techniques and procedures, which are used to work upon the images of documents and exchange them from pixel information into a format that can be read by a computer. Information retrieval from the document images is a challenging task, hence number of techniques and procedures are used for document image processing [3]. Converting a scanned grey scale image into a binary image, the foreground (or regions of interest) and removing the background is an important step in many image analysis systems includes document image processing. Historical documents are also prone to being attacked by pests and insect's image connecting past and present is essential in order for one to find the right path towards future [5]. It refers to digital images of symbolic objects such as scanned documents, postal addresses, printed articles, forms, engineering drawings, topographic maps, license plates, billboards, subtitles in photos and video. Scanners, Printers, Fax machines are the source for these images. Document image analysis deals with solutions to obtain computer readable description from document images. Recognition and extraction of text and graphics components for various applications is the aim of document image analysis [11]. The document image retrieval is concerned with content based document browsing, indexing and searching [15] from database of document images.

II. DOCUMENT IMAGE ANALYSIS

The document image analysis identifies the text and graphics components in images of documents, and to extract the feature information. Two different categories of document image analysis, they are Text and Graphics processing. Figure 1 depicts the document image processing.

A. Text Processing

1) It compacts with the textual components of a document image
2) Determining the skew (the document scanned in the computer used document image).
3) Finding columns, paragraphs, textual lines, words, recognizing the text by using OCR.

B. Graphical Processing

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

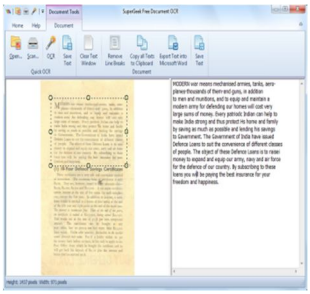
©IJRASET (UGC Approved Journal): All Rights are Reserved

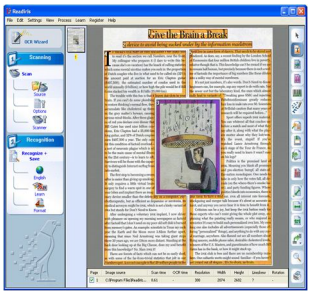
International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

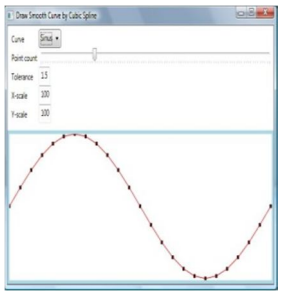
Document image

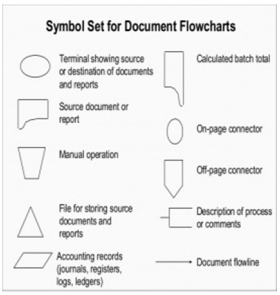
Textual Processing

Graphical Processing









OCR
Page Layout Analysis
Line Processing
Region and Symbol

Text Document

Skew, Text lines, Text blocks, and Paragraphs

Straight lines, Corners and Curves

Filled Regions

Figure1 Document image analysis using text processing and graphical processing.

Sample document images for textual processing and graphical processing.

Figure 2 Textual processing

Figure 3 Graphical processing

©IJRASET (UGC Approved Journal): All Rights are Reserved

III.APPLICATIONS AREAS OF DOCUMENT IMAGE ANALYSIS

The document image analysis and retrieval has lot of application areas. Some of the important applications are listed in given below:

1) Signature verification is used in bank sector.
2) In forensic offices and crime department's it is used for detecting forgery.
3) Reading and recognizing number plates of vehicles using the database stored in the computers during the time of vehicle registration.
4) Tracking vehicles to detect over speed and breaking of any traffic rules.
5) Arrangement of large document datasets such as legal, historical or security related documents.
6) Document image analysis can be used for designing better search engines on the Web.
7) Name, Address, Location and Pin code detail extraction from the mails in couriers and in postal department.
8) Document image analysis can also be used for compression of image documents.
9) Identification of scripts in different languages is also possible with the help of Document Image Analysis [11]
10) *Title based searching:* This application helps the users to search the particular word in Document images. Document similarity measurement: It allows the user to retrieve documents by specifying complete document image as a query rather than a keyword
11) *Document image retrieval using signature as queries:* This application includes retrieving Documents from the database by specifying the document containing signature as a query. Here the signature features from the query document are used for retrieval of documents
12) *Logo based document retrieval:* Such applications use a document containing logo as a Query to retrieve all documents from the database that contain a logo similar to that of query Document. *Retrieving imaged documents in digital libraries:* The document image retrieval helps Users to search particular keywords, titles, subtitles and images from a list of articles that are Stored in digital libraries.

IV. DOCUMENT IMAGE RETRIEVAL

Document image retrieval is a document management and document analysis problems such as page segmentation and OCR. The can achieve excellent performance in dealing with clean document image.

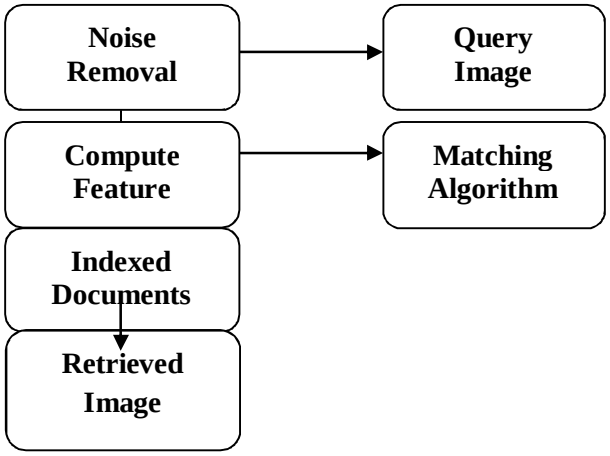


Figure 4 Document Image Retrieval

A. Noise Removal

The noise removal Dots can be modeled as impulses (salt-and-pepper or speckle) or continuously varying (Gaussian noise).The different types of noise removal using document image analysis Amplifier Noise (Gaussian noise), Salt-and-Pepper Noise (Impulse Noise), Shot Noise, Quantization Noise (Uniform Noise), Film Grain, Non-Isotropic Noise, Speckle Noise (Multiplicative Noise), Periodic Noise. After the detection of marginal noise regions, different removal methods are performed. The arisen impulse noises display as light and dark noise pixels under random distribution on the image. For an image corrupted by noises, we can use linear or nonlinear filter methods to reduce noises. Spatial and frequency domain algorithms based on type of the noise in document can be employed. Marginal noise detection will reduce an original document image into a smaller image, and then find marginal noise regions according to the shape length and location of the split blocks. After the detection of marginal noise regions, different

©IJRASET (UGC Approved Journal): All Rights are Reserved

removal methods are performed. A local thresholding method is proposed for the removal of marginal noise in gray-scale document images, whereas a region growing method is devised for binary document images noise various types of noise Gaussian noise, Poisson noise, Speckle noise a d so on.

B. Feature Extraction

Feature extraction is a type of dimensionality reduction, which efficiently represents parts of an image as a compact feature vector. This approach is useful when image sizes are large and a reduced feature representation is required to quickly complete tasks such as image matching and retrieval. Instead of extracting features of document images existing in the database every time during retrieval, they are extracted and stored only once

C. Similarity Matching

Similarity matching or generally image matching two images are compared to determine their visual similarity and whether they are equivalent (i.e., scanned from the same original). Matching algorithm used to compare features of query image with the Indexed features of the images present in the database of documents. To measure the similarity using ranked based on the distance value. The Compare these image matching algorithm on the basis of various measures such as accuracy, processing speed, flexibility to use for various data sets, invariance to rotation, scale and illumination.

D. Ranking of the Documents

The results of the matching algorithm to ranked in increasing order of the similarity distance. This step aim at organizing retrieved document as to find closest document to the query and other documents that are nearly matching with the query [4].The many different techniques to find out the hidden patterns like text mining, information Extraction, information Retrieval. Information retrieval models are used for ranking relevant document.

V. ISSUES IN DOCUMENT IMAGE PROCESSING

A. Computational speed

Step by step approach is required for Document image analysis and retrieval. These steps may include document capturing, feature extraction, feature analysis, matching the features, ranking of documents. These steps are computationally expensive. Hence there is need for optimization of these operations during retrieval to satisfy the need of the users in practical applications.

B. Degradation of documents

The degradation of printed or scanned document images can be due to several reasons. Some of the reasons are excessive dusty noise, large ink-blobs joining disjoint characters or components, poor quality of paper and ink, text overlapping the signature. Image enhancement or noise reduction techniques need to be developed especially for improving quality of degraded document images before processing.

C. *Language dependency*

The character shapes, orientation and methods used for representation of documents vary from language to language. Thus it poses a challenge for the researcher's to develop and implement language independent document analysis and retrieval algorithms.

D. Standardization of datasets

The availability of standard dataset for performance evaluation is another issue for DIAR research community. Almost each paper deals with different datasets and this may be due to the variety of problems addressed by different applications. Development of common platform for testing the document image analysis/retrieval algorithms and techniques is an important issue yet to be addressed.

VI. CONCLUSION

Document image analysis is one of the important and challenging research domains in the field of computer science. Most of the researchers are interested to do their research work in the field of document image processing. Many challenging research problems are available in document image analysis. These problems can be solved by developing new algorithms, concepts and techniques Document image basic concepts, essential characteristics, and their applications. Emerging and open research issues in document image analysis also described in this paper.

©IJRASET (UGC Approved Journal): All Rights are Reserved

REFERENCES

[1] L.O. Gorman and R. Kasturi, "Document image analysis: An executive briefing", 1999.
[2] Lawrence O'Gorman "Document Image Analysis" IEEE Computer Society Executive Briefings, ISBN 0-8186-7802-X Library of Congress Number 97-172831997.
[3] Nawei Chen · Dorothea Blostein"A survey of document image classification: problem statement,classifier architecture and performance evaluation"1 June 2004.
[4] Kiranpreet and Ramandeep kaur, "Survey On Document Image Processing"International Journal of Computer Trends and Technology (IJCTT) – volume 15 number 1 – Sep 2014.ISSN: 2231-2803
[5] Umesh D. Dixit1 and M. S. Shirdhonkar, "Document Image Analysis and Retrieval System", International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 2, April 2015, DOI: 10.5121/ijci.2015.4225 259"
[6] Er. Varun Kumar, Ms. Navdeep Kaur, and Er. Vikas, "A Survey on Various Approaches of Degraded Document Image Enhancement", International Journal for Research in Applied Science &EngineeringTechnology (IJRASET).
[7] K.Ajitha, "A Survey on Degraded Document Image Binarization Techniques" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 11, November 2014 ISSN: 2278 – 1323.
[8] Greeshmamol Varghese and Kumudha Raimond, "A Survey on the Methods Used in Document Digitization and its Applications", International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3, Issue 12, December 2013 ISSN: 2277 128X
[9] S. Shukla, "Survey on Image Mining, its Techniques and Application", International Journal of Computer Application, Volume 133 – No.9, January 2016, (0975 – 8887)
[10] Ying Liu and Sargur N. Srihari, "Document Image Binarization Based on Texture Features", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 5, May 1997.
[11] D. Lopresti and J. Zhou, "Document analysis and the world wide web", IAPR Workshop Document Analysis Systems, 1996, pp.651–669
[12] Y.He, Z. Jiang, B. Liu, and H. Zhao, "Content-Based Indexing and Retrieval Method of Chinese Document Images," In Proc. Fifth Int'l Conf. Document Analysis and Recognition (ICDAR'99), pp.685-688, 1999.
[13] D.Niyogi and S. Srihari, "The Use of Document Structure Analysis to Retrieve Information fromDocuments in Digital Libraries," In Proc. SPIE, Document Recognition IV, vol. 3027, pp.207-218,1997.
[14] Sanjay T.Gandhe, K. T. Talele and Avinash G. Keskar, "Image Mining Using Wavelet Transform", Knowledge-Based Intelligent Information and Engineering Systems, Springerlink book chapter, pp. 797-803, 2007.
[15] Harini. D. N. D and Dr. Lalitha Bhaskari. D, "Image Mining Issues and Methods Related to Image Retrieval System", International Journal of Advanced Research in Computer Science, Volume 2, No. 4, 2011.
[16] Janani M and Dr. ManickaChezian. R, "A Survey On Content Based Image Retrieval System", International Journal of Advanced Research in Computer Engineering & Technology, Volume 1, Issue5, pp 266, July 2012.
[17] G. Nagy, "20 Years of Document Image Analysis in PAMI", IEEE Transactions on PatternAnalysis and Machine Intelligence, Vol. 22, No. 1, January 2000.

©IJRASET (UGC Approved Journal): All Rights are Reserved