Document Image Retrieval: An Overview

Manesh B. Kokare
S. G.G.S
Institute of Engineering and Technology,
Nanded-431606, India

M.S.Shirdhonkar
B.L.D.E.A's
College of Engineering
and Technology, Bijapur-586103, India

ABSTRACT
The economic feasibility of creating a large database of document image has left a tremendous need for robust ways to access the information. Printed documents are scanned for archiving or in an attempt to move towards a paperless office and are stored as images. In this paper, we provide a survey of methods developed by researchers to access document images. The survey includes papers covering the current state of art on the research in document image retrieval based on images such as signature, logo, machine-print, different fonts etc.

Categories and Subject Descriptors
D.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; D.2.8 [**Database Management**]: Database Applications-*image database*;

General Terms
Algorithm, Design, Experimentation, Theory.

Keywords
Document image retrieval, Query image based-document retrieval, Information Retrieval, Word searching.

1.1INTRODUCTION
Document image retrieval is a very attractive field of research with the continuous growth of interest and increasing security requirements for the development of the modern society. Good documents essentially play an important role in our day to day life. Complex documents present a great challenge to the field of document recognition and retrieval. The combined presence of noise, handwriting, signature ,logos, machine-print with different fonts and rule lines impose a lot of restrictions to algorithms that work relatively well on simple documents.

The primary task of processing these complex documents is to isolate the different contents present in the documents. Once the contents are separated out, then they can now be called as indexed documents which are ready to use for a content-based image retrieval system. The document image understanding , covering a variety of documents such as bank checks [1], business letters [2], forms [3], and technical articles [4-5], has been an interesting research area for a long time. In the context of document image retrieval, logo provides an important form of indexing that enable effective explanation of data [6].Given a large collection of documents, searching for a specific logo is a highly effective way of retrieving documents from the associated organization. Building an effective access to these document images requires designing a mechanism for effective search and retrieval of image data from document image collection. In searching complex documents, such as repository of archival office documents, a task of relevance is relating the signature in a given document to the closest matches within a database of documents; this is known as signature retrieval task. Given a database of signed document, it would be of interest to relate a queried document to other documents in this database which have been signed by the same author.

The main contributions of this paper are summarized as follows. Firstly, in this paper, we have provided detailed survey of document image retrieval. Secondly, we have discussed about the applications and challenges in document image retrieval. Finally, future directions in document image retrieval are also suggested.

The paper is organized as follows .In section 2, we discuss the system architecture and in section 3, we discuss applications. In section 4, we review the current state of art on document image retrieval research activities. The challenges in the section 5, deal with challenges in the field of document image retrieval. In section 6, we discuss the evaluation strategy. Section 7, concludes the paper.

2.THE SYSTEM ARCHITECTURE
A block diagram shown in figure1, describes the steps involved in document image retrieval System (DIRS). The various steps involved in document image retrieval are noise removal, feature extraction, and matching algorithm, which are discussed here.

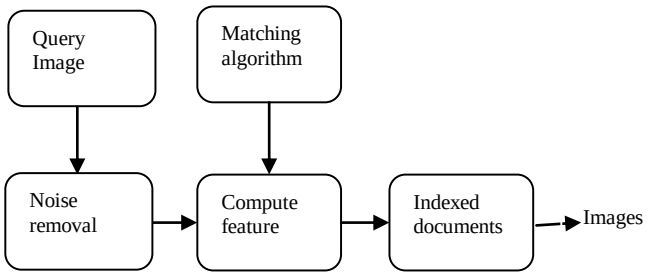

Figure 1: A Block diagram describing the steps involved in document image retrieval system.

2.1Noise removal

Digital capture of images can introduce noise from scanning devices and transmission media. Noise removal is carried out to get rid of any noise or printed text overlapping the extracted images such as signature, logos, machine-print etc. In the pre-processing step the printed text is removed from the image samples. To remove the printed text from images variety of methods can be used such as image enhancement procedures based on chain code [7], Support Vector Machine (SVM) [8], to classify each connected component as a part of noise components, signature, a small handwritten text, logo, noise etc.

2.2Feature extraction

Feature extraction involves extracting the meaningful information from the document images. So that it reduces the storage required and hence the system becomes faster and effective in document image retrieval. Once the features are extracted, they are stored in the database for future use. The degree to which a computer can extract meaningful information from the image is the most powerful key to the advancement of intelligent image interpreting systems. One of the biggest advantages of feature extraction is that, it significantly reduces the information (compared to the original image) to represent an image for understanding the content of that image. It uses various techniques to extract the features such as Gradient, Structural and Concavity (GSC) features [9-10], which measures the image characteristics at local, intermediate and large scale, features based on density distribution and key block features [11], fisher classifier [6], Angular radial partitioning of a images regions [12], Conditional Random Field [13], DTW [14], etc. are used for feature extraction.

2.3Matching algorithm

The document image retrieval is performed using a matching algorithm to compare the query image with image database. Figure 1, Shows the various operational steps in the retrieval process:

Step 1: Noise removal from the query image;

Step 2: Feature extraction from the query after noise removal;

Step 3: Matching the query image features to each of the indexed documents ;

Step 4: Ranking the documents in accordance with the results from the matching algorithm

The task of matching algorithm is to compare this feature with the indexed features of the image present in the database of documents. Similarity measure, the query image feature vector and database image feature vector is compared using the distance metric [15]. The images are ranked based on the distance value. The comparison of different metrics such as Manhattan, Euclidean, Chebychev, etc. is done in [16]. The normalized correlation is considered to be better for binary feature vectors as characterization to other metrics.

3.APPLICATION

Searching and retrieval of documents has been a topic of interest for many years. There are various applications of document image retrieval such as,

3.1 Word searching

Searching / locating a user-specified keyword in image format documents have been of interest. It has its practical value for document information retrieval. The users can locate a specified word in document images without any prior need for the images to be OCR-processed [18].

3.2 Document similarity measurement

Measuring the similarity between documents has practical applications in document image retrieval. For instance, a user may use an entire document rather than a keyword to retrieve documents whose content is similar to the queried document [18].

3.3 Document image retrieval using signature as queries

In searching complex documents, such as a repository of archival office documents, a task of relevance is, relating the signature in a given document to the closest matches within a database of documents; this is known as the signature retrieval task. For a given database of signed documents, it would be of interest to relate a queried document to other documents in this database which have been signed by the same author [7].

3.4Automatic document logo detection

Logos are commonly used in business and government documents as a declaration of document source and ownership. In the context of document image retrieval, logos provide an important form of indexing that enables effective exploration of data. Given a large collection of documents, searching for a specific logo is a highly effective way of retrieving documents from the associated organization [6].

3.5Retrieving imaged documents in digital libraries

A great number of documents are scanned and archived in the form of digital images in digital libraries, to make them available and accessible in the internet. By making digital library an online store of books, magazine, student thesis, etc, with the help of document image retrieval system, users would be able to do a search on a set of keywords, and get a list of relevant article, for viewing or printing [20].

4.RELATED WORK

A number of attempts have been made to bring the document analysis and information retrieval communities together to work on the retrieval of noisy data [19] from 1992 through 1996, the University of Nevada, Las Vegas, held an annual Symposium on Document Analysis and Information retrieval [20]. The goal of the Symposium was to offer a strong program of basic research in the complementary fields of document analysis and information retrieval .In 1994, Tang et al. [21], proposed methods for automatic knowledge acquisition in document images by analyzing the geometric structure and logical structure of the

images. In 1997, Niyogi et al. [18], described an approach to retrieve information from document images stored in digital library by means of knowledge-based layout analysis and logical structure derivation techniques, in which significant sections of documents, such as the title, are utilized. In 2000, Liu et al. [22], presented an approach to image-based form document retrieval. They proposed a similarity measure for forms that is insensitive to translation, scaling, moderate skew, and image quality fluctuations, and developed a prototype form retrieval system based on their proposed similarity measure. In the domain of Chinese document image retrieval method based on the stroke density of Chinese characters [23]. This method enables fast retrieval of Chinese printed document image using the index method. This method does not support for font style changes provision of document images retrieval. In 2003, Chalechae, et .al [12], proposed signature based decomposition and retrieval of document images. As a case study he investigated, Arabic/Persian signature recognition and retrieval. Connected component analysis and labeling along with geometric properties are used to recognize the signature region. An angular radial partitioning scheme is then introduced for the description of spatial distribution of pixels in the interested region. This method shows the supremacy of the retrieval performance of proposed approach over the line segment distribution method. In 2004, Liz hangs, et. al [24], proposed for designing an information retrieval system with ability of dealing with imaged document stored in digital libraries .The proposed system provide an efficient and promising tool for document image retrieval .This method is used to propose word coding techniques .This method does not support to integrate linguistical knowledge to present system.

In 2005, Liu, et .al [11], presented a method of document image retrieval based on density distribution features and key block feature of document image. Key block features are applied to confirm the reliability of the raw candidate's images so as to improve the retrieval performance. This method is suitable for retrieving color and gray document images from a large scale hybrid document image database of different languages in real time. Nakai, et .al [25], proposed a method of camera based document image retrieval which is characterized by indexing with geometric invariants and voting with hash tables. This method shows high accuracy with normal digital camera. In 2006, Bal Subramanian, et .al [14-15], proposed a system for retrieval of relevant documents from large document image collections. It achieves effective search and retrieval from a large collection of printed document images by matching image features at word level. For representation of the words Dynamic Time Wrapping (DTW) based features are employed. Srihari, et. al [26], proposed a document image retrieval using signature as queries. A signature retrieval strategy includes techniques for noise and printed text removal from signature images. In this method signature matching is based on a normalized correlation, similarity measure using global shape based binary feature vectors.

In 2007, Schomaker [27], proposed retrieval of handwritten lines of text in historical documents .This method used brute-force matching line-strip images using a correlator and results are compared to feature based methods. Zhu et. al [6], proposed automatic document logo detection and extraction in document images that robustly classifies and precisely localizes logos using a boosting strategy across multiple image scales. Spitz described character shape codes for duplicate document detection [28], information retrieval [29], word recognition [32-33], and document reconstruction [31], without resorting to character recognition. Character shape codes encode whether or not the character in question fits between the baseline and the x-line or, if not, whether it has an ascender or descender, and the number and spatial distribution of the connected components. To get character shape code, character cells must first be segmented. This method is therefore unsuitable for dealing with words with connected characters. Additionally, it is lexicon-dependent, and its performance is somewhat affected by the appropriateness of the lexicon to the document being processed. In 2007, Guillaume Joutel, et .al [39], proposed, curvelets based feature extraction of handwritten shapes for ancient manuscripts classification. This approach is language independent, visual orientation and appearance based document image retrieval.

In 2008, Shijian Lu, et.al [32], proposed, document image retrieval through word shape coding. It retrieves document image by a new word shape coding scheme, which captures the document content through annotating each word image by a word shape code. We annotate word image by a set of topological shape features including character ascenders/descanters, character holes and character water reservoirs. This approach shows that document image retrieval technique is fast, efficient and tolerant to various types of document degradation. Guangya Zhu, et. al [40], proposed, a signature-based document image retrieval system .It automatically detects, segments, and matches signatures from document images with unconstrained layout and complex backgrounds.

In 2009, Guangya Zhu, et.al [41], proposed, an automatic logo-based document image retrieval system. In this approach used logo detection, segmentation by boosting a cascade of classifiers across multiple image scales and logo matching using translation, scale and rotation invariants of shape descriptors. This method is segmentation free and layout independent. Ehtesham Hassan, et.al [42], proposed a shape descriptor based document image indexing and symbol recognition. In this method hierarchical distance based hashing technique is used for document image indexing. Tilin Li, et.al [43], proposed system for document image retrieval with local feature sequences. This method presents a fast, accurate and OCR-free image retrieval algorithm using local features and intrinsic, unique , page-layout free characteristics of document images.

5.CHALLENGES IN DESIGN AND IMPLEMENTATION OF THE DIRS

Complex documents pose a great challenge in the field of document recognition and retrieval. Search and retrieval from large collection of document images is one of the important issues [14].To design and implement a successful search engine in image domain, we need to address the following open issues.

5.1Computational speed

Searching from large collection of document images passes through many steps: Image processing, feature extraction,

matching and retrieval of documents. Each of these steps could be computationally expensive. Hence there is need for optimal use of operations during retrieval.

5.2Degradation of documents

The degradation of the printed document image can be due to several reasons:

- Excessive dusty noise, logos, figures, printed and Handwritten text etc.
- Large ink-blobs joining disjoint characters or components
- Degradation of printed text due to the poor quality of paper and ink.
- Text overlapping the signature

The design of an appropriate representation of scheme and matching algorithms to accommodate the effect of degradation is necessary.

5.3Need for cross-lingual retrieval

The documents that users need may be available in different languages. Most educated Indians can read more than one language. Hence, we need to design a mechanism that allows users to retrieve all documents related to their queries in any of the Indian languages.

5.4Indian language

Indian languages pose many additional challenges. Some of these are

- Lack of standard representation for the fonts and encoding.
- Lack of support from operating system, browser, and keywords.
- Lack of language processing routines.

These issues add to the complexity of the design and implementation of a document image retrieval system.

6.EVALUATION STRATEGIES

In document image retrieval system, a strategy for evaluation involves determining the following aspects [38].

6.1An appropriate database for evaluation

The dataset should be general enough to cover a large range of semantics from a human point-of-view and the evaluation to be statistically significant.

6.2An appropriate metric and criteria for evaluating competing approaches

The evaluation criteria should try to model human requirements from a population perspective. Evaluation metrics have been quite naturally adopted from document image retrieval research. Two of most popular evaluation measures are described as follows.

- Precision: This refers to the percentage of retrieved document images that are relevant to the query

$$precision(N) = \frac{R_n}{N} \quad (1)$$

Where N is number of retrieval and R_n is number of relevant matches among retrievals.

- Recall: This pertains to the percentage of all the relevant document images in the search database which are retrieved.

$$\text{Re } call(N) = \frac{R_n}{M} \quad (2)$$

Where M is the total no. of relevant matches in the database R_n is number of relevant matches among retrievals.

7.DISCUSSION AND CONCLUSION

Today information technology has proved that there is a need to store, query, search and retrieve large amount of electronic information efficiently and accurately. So document image retrieval is very challenging field of research with the continuous growth of interest and increasing security requirements for the development of the modern society.

This paper surveys the technical achievements in the field of document image retrieval, discuses system architecture, comprehensive survey of various proposed methods to retrieve the documents. It also highlights the challenges and scope of research

8.REFERENCES

[1] S.Djeziri, F.Noubound, and R.Plamondon.1998.Extraction of Signature from Check background Based on A Filitormity Criterion. IEEE. Trans. Image processing, vol.7. no.10, pp.1424-1438.

[2] A. Dengel.1993.Initial of Document Structure. In Proc. IEEE Second Inf. Conf. Document Analysis and Recognition, pp.86-90.

[3] B.Yu and A.K.Jain.1996. A Generic System for Form Dropout. IEEE Trans. Patt. and Mach. Intell, vol.18, no.11, pp1127-1134.

[4] A.K. Jain and B. Yu.1998.Document Representation and Its Application to Page deposition. IEEE Trans. And Mach. Intell., vol.20, no.3, pp.294- 308.

[5] G. Nagy, S.Seth and M. Viswanathan.1992. A Prototype Document Image Analysis System for Technical Journals. IEEE Computer, vol.25, no.7, pp.10-22.

[6] Guangyu Zhu and David Doermann. Automatic Document Logo Detection. Institute for advanced computer studies, University of Maryland, College Park, MD 20742, USA

[7] S. Srihari, S. Shetty, S. Chen, H. Srinivasan and C. Huang.2006.Document Image Retrieval using Signatures as Queries. In Proceeding of the Second International Conference on Document Image Analysis for libraries (DIAL'06).

[8] C. Barges.1999.A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2):121-167.

[9] J.T.Favata and G. Srikantan.1996.A Multiple Feature Resolution Approach for Hand/ Printed Digit and Character Recognition. International Journal of Imaging Systems and Technology, 7:304-311.

[10] G. Srikantan, S. Lam, and S.Srihari.1996.Gradient Based Contour Encoding for Character Recognition. Pattern Recognition, 29(7):1147-1160.

[11] Hong Liu, Suoqian Feng, Hong bin Zha, Xueping Liu.2005. Document Image Retrieval Based On Density Distribution Feature and Key Block Feature. In Proceeding of 2005 conference on Document Analysis and Recognition (ICDAR'05).

[12] Abdullah Chalechale, Golshah Naghdy. Signature Based Document Retrieval. Faculty of information-papers, University of Wollongong.

[13] Shravya Shetty ,Harish Srinivasan , Matthew Beal and Sargur Srihari, "Segmentation and Labeling of Documents using Conditional Random Fields", Center of Excellence for Document Analysis and Recognition (CEDAR) , University of Buffalo, State University of New York.

[14] A. Balasubramanian, Million Meshesha and C.V. Jawahar.2006. Retrieval from Document Image Collections", Springer -Verlag Berlin Heidelberg.

[15] C.V.Jawahar, Million Meshesha, A. Balsubramanium.2004.Searching in Document Images.In Proc. of the 4th Indian Conference on Computer Vision, Graphics and Image Processing, pp.622-627.

[16] Manesh Kokare, B.N.Chatterji and P.K.Biswas.2003. Comparison of Similarity Metrics for Texture Image Retrieval. In IEEE TENCON.

[17] Ritendra Datta, Dhiraj Joshi, Jiali, and James Z. Wang.2008. Image Retrieval: Ideas, influences, and Trends of the New Age. ACM Computing Surveys, vol. 40, no.2.

[18] D.Niyogi and S. Srihari.1997.The Use of Document Structure Analysis to Retrieve Information from Documents in Digital Libraries. In Proc. SPIE, Document Recognition IV, vol.3027, pp.207-218.

[19] Doermann, D.1998.The Indexing and Retrieval of Document Images: A Survey. Computer Vision and Image Understanding (CVIU) 70, pp.287-298.

[20] Symposium on Document Analysis and Information Retrieval.1992 and 1993. University of Nevada, Las Veg s,

[21] Y.Tang, C.D.Ya, and C.Y.Suen.1994.Document Processing for Automatic Knowledge Acquisition. EEE Trans. Knowledge and Data Eng., vol.6, no.1, pp.3-21.

[22] J. Liu and A.K.Jain.2000.Imaged-Based Form Document Retrieval.Pattern Recognition, vol.33, no.3, pp.503-513.

[23] Y.He, Z. Jiang, B. Liu, and H. Zhao.1999.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.

[24] Yue Lu and Chew Lim Tan.2004.Information Retrieval in Document Image Databases. IEEE Tran. on Knowledge and Data Eng., vol.16, no. 11.

[25] Tanohiro Nakai, Koichi Kise, Masakazu Iwamura.2005. Camera Based Document Image Retrieval as Voting for Partial Signatures of Projective Invariants. Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition.

[26] Sargur N. Srihari, Shravya Shetty, Gady Agam and Ophir Frieder.2006. Document Image Retrieval Using Signature as Queries. In Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06).

[27] L.R.B. Schomaker. Retrieval of Handwritten Lines in Historical Documents.

[28] A.L.Spitz.1997. Duplicate Document Detection. in Proc. SPIE, Document Recognition IV, vol.3027, pp.88-94.

[29] A.F. Smeaton and A. I. Spitz.1997.Using Character Shape Coding for Information Retrieval. In Proc. Fourth Int'l Conf. Document Analysis and Recognition, pp.974 –978.

[30] A.L.Spitz.1999.Shape Based Word Recognition. Int' J.Document Analysis and Recognition. vol.1, no. 4, pp.178-190.

[31] A.L.Spitz.2002.Progress in Document Reconstruction. In Proc. 16th Int'l Conf. Pattern Recognition, vol.1, pp. 464-467.

[32] Shijian Lu, Linlin Li and Chew Lim Tan.2008.Document Image Retrieval through Word Shape Coding. IEEE Trans. Patt. And Mach. Intell., vol.30, no.11, pp1913-1918.

[33] S.Lu and C.L. Tan.2008.Retrieval of Machine-Printed Latin Documents through Word Shape coding. Pattern Recognition, vol.41, no.5, pp.1816-1826.

[34] B.Zhang, S.N. Srihari, and C.Hung. 2004. Word Image Retrieval Using Binary Features. Document Recognition and Retrieval XI, SPIE, San Jose, CA.

[35] Digital Image using MATLAB by Rafael C. Gonzales, Richard E. Words and Steven L Eddins, Low Price Edition, India.

[36] A. K. Jain, Fundamentals of Digital Image Processing, Prentice Hall of, India.

[37] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer.

[38] Ritendra Datta, Dhiraj Joshi, Jiali, and James Z. Wang.2008.Image Retrieval: Ideas, influences, and Trends of the New Age. ACM Computing Surveys, vol. 40, no.2, article 5.

[39] Guillaume Joutel, Veronique Eglin, Stephane Bres, and Hubert Emptoz.2007.Curvelets based features extraction of handwritten shapes for ancient manuscripts classification. Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol.6500, 65000D-1-11.

[40] Guangyu Zhu, Yefeg Zheng, and David Doermann.2008.Signature based document image Retrieval. ECCV, 2008, Part III, LNCS 5304, pp.752-765.

[41] Guangyu Zhu and David Doermann.2009.Logo detection for document image retrieval. 10th International Conference on Document Analysis and Recognition, pp.606-610.

[42] Ehtesham Hassan, Santanu Chaudhury, and M Gopal.2009.Shape descriptor based document image indexing and symbol recognition. 10th International Conference on Document Analysis and Recognition, pp.206-210.

[43] Jilin Li, Zhi-Gang Fan, Yadong Wu and Ning Le.2009. Document image retrieval with local features sequences. 10th International Conference on Document Analysis and Recognition, pp.346-350.