A SURVEY ON DOCUMENT IMAGE ANALYSIS AND RETRIEVAL SYSTEM

Umesh D. Dixit[1] and M. S. Shirdhonkar[2]

[1]Department of Electronics & Communication Engineering, B.L.D.E.A's CET, Bijapur.
[2]Department of Computer Science and Engineering, B.L.D.E.A's CET, Bijapur.

ABSTRACT

The digitization of documents and their availability over the network demands solution toward content based document image analysis, indexing, searching and retrieval. Signature, Logo and Layout of the documents present convincing evidence and provide an important form of indexing for effective document image retrieval in a variety of applications. This paper describes methods and techniques developed for document image analysis and retrieval by researchers.

KEYWORDS

Feature extraction, Image analysis, Document retrieval, Pre-processing, Document indexing

1. INTRODUCTION

Document image 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. The aim of Document Image Analysis and Retrieval System (DIARS) is finally to develop paperless office.

Document image analysis is subfield of document image processing. 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]. However document image retrieval is concerned with content based document browsing, indexing and searching [30] from huge database of document images.

Complex documents present a great challenge to the field of document image analysis and retrieval. The primary task of analyzing these complex documents is to isolate the different contents present in the documents such as graphical and textual components. Once the contents are separated out, they can be indexed and used for content-based image retrieval. The document image recognition and understanding, covering a variety of documents such as bank cheques, business letters , different forms and technical articles, have been an interesting area of research for a long time. Signature, logo, handwritten text and layouts provide a convincing form of indexing and enable effective description of data [19] for document retrieval.

The main contribution of this paper is:

- It provides detailed survey of document image analysis, retrieval methods and techniques.

DOI: 10.5121/ijci.2015.4225

- Highlights the classification, applications, challenges and issues in the area of document image analysis and retrieval.

This paper is organized as follows: section 2 provides a general framework for document image analysis and retrieval, section 3 will review the current state of art on document image analysis and retrieval research activities, section 4 provides an insight to Document Image Retrieval classification, section 5 discuss the applications. In section 6 we discuss the issues and challenges, in section 7 evaluation strategies and section 8 concludes the paper.

2. GENERAL FRAME WORK FOR DIARS

General framework for document image analysis and document image retrieval is described in this section.

2.1 Document image analysis

Figure 1 shows the sequence of steps for document image analysis [11]. The steps include capturing data from document, Binarisation, Feature level analysis, Text and graphical components' analysis and recognition.

- **Data capturing:** The data in paper document are usually captured using a scanner and stored in a file in the form of image with an extension jpeg, tiff, bmp, png etc. The captured document may be a colour or gray scale image.

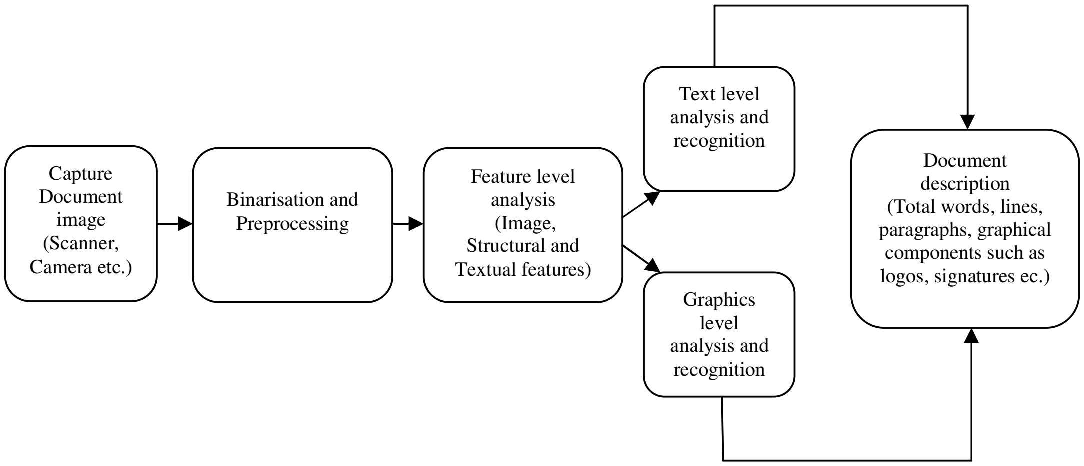
Capture Document image (Scanner, Camera etc.)
Binarisation and Preprocessing
Feature level analysis (Image, Structural and Textual features)
Text level analysis and recognition
Graphics level analysis and recognition
Document description (Total words, lines, paragraphs, graphical components such as logos, signatures ec.)

Figure1. Steps involved in document analysis

- **Binarization and preprocessing:** To separate foreground and background information the document image is binarized. This step converts the captured document into pixels with intensity level 1 or 0. Preprocessing includes noise reduction, segmentation and converting the image into required form for further processing. Noise in the document image is possible from many sources including degradation due to aging, photocopying, or during data capture. Segmentation process is carried out to separate textual and graphical components of the document. Segmentation of textual information helps in locating columns, paragraphs, words, and characters and segmentation of graphics is to separate out symbols, logo, signature, lines from the document image.

- **Feature level analysis:** This step is to carry out analysis of textual and graphical components. Image features, Structural features or Textual features are extracted from the document image. These features may be either local or global features. The table 1 summarizes the different features used for document image analysis.

Table 1. Different features used in Document Image Analysis

Image features	Structural features	Textual features
1. Connected components. 2. Gaps between row/Columns. 3. Location and size of cells. 4. Text histogram.	1.Physical Layout. 2.Logical structures 3.Results of functional labeling 4.Spatial relations	1. Page/Text layout features. 2. Textual features from OCR results analysis

- **Text level analysis and recognition:** Two main types of analysis are applied to text in documents. First one is Optical character recognition (OCR) to extract the meaning of the characters and words from document images. The second method is page-layout analysis to recognize formatting of the text in document image that includes text bodies in different functional blocks, headers, footers, titles, subtitles etc.

- **Graphics level analysis and recognition:** Graphics level analysis includes recognizing lines, curves and other features to identify different components such as signatures, logos and layout of the document for further inspection.

- **Document description:** The result of document image analysis is document description and consists of both textual and graphical components present in the document.

2.2 Document image retrieval

Figure 2 shows the steps involved in document image retrieval. These steps are briefly explained below.

- **Noise removal:** When an image is captured through a digital device it is very natural that a noise may get introduced with required information. Image enhancement techniques can be applied to reduce this noise. Spatial and frequency domain algorithms based on type of the noise in document can be employed.

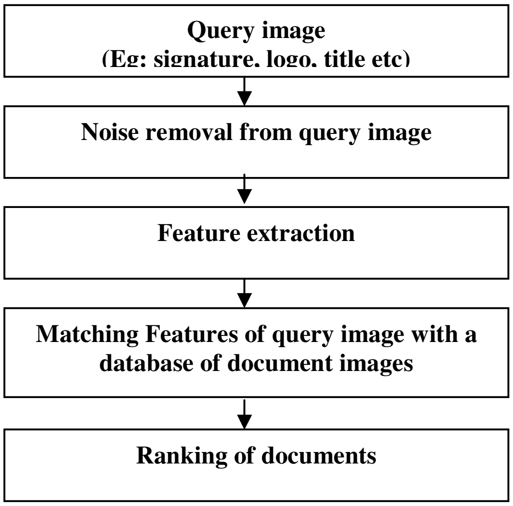
Query image
(Eg: signature, logo, title etc)
Noise removal from query image
Feature extraction
Matching Features of query image with a database of document images
Ranking of documents

Figure 2. Steps involved in document image retrieval

- **Feature extraction:** Meaningful information is extracted from the document images at this stage for document image retrieval. The extracted features may decide the efficiency analysis and retrieval. Instead of extracting features of document images existing in the database every time during retrieval, they can be extracted and stored only once. This makes the system faster and effective for document image retrieval. The different features like Gradient, Structural and Concavity (GSC) can be extracted from the document image for implementing document image retrieval. Density distribution and key block features [14], fisher classifier [19], Angular radial partitioning of image regions [12], Conditional Random Field [3], DTW [16] can also be considered.
- **Similarity matching:** Matching algorithm is used to compare features of query image with the indexed features of the images present in the database of documents. To measure the similarity, the query image feature vector and database image feature vector are compared using the distance metric such as Manhattan, Euclidean, Chebychev etc. The images are then ranked based on the distance value.
- **Ranking of the documents:** The results of the matching algorithm are ranked in increasing order of the similarity distance. This step aims at organizing retrieved documents so as to find closest document to the query and also other documents that are nearly matching with the query document.

3. RELATED WORK

Lot of methods and techniques are evolved for document image analysis and retrieval. This section briefly describes and highlights the work done in this area by researchers.

Lot of efforts are made to bring the document analysis and information communities to work together on retrieval of noisy data from document images [7]. In 1994, Tang et. al [1] presented techniques for extracting knowledge from document images. They employed geometric structure and logical structure to acquire this knowledge. Niyogi and Srihari [4] proposed a method for retrieving information from digital libraries. Section of documents like titles are used as query for retrieval of data with the help of knowledge-based layout analysis and logical structure derivation techniques.

In 2000, Liu and Jain [10] presented an approach to image-based form document retrieval. They proposed a prototype form retrieval system using similarity measure for forms that is insensitive to translation, scaling, moderate skew, and image quality fluctuations. In the domain of Chinese document image retrieval, a method based on the stroke density of Chinese characters [9] is proposed. They used index method for retrieval of Chinese printed document images relatively faster. However this method does not support for change in font style. In 2003, Chalechae and Naghdy [12] proposed signature based decomposition and retrieval of document images. As a case study they investigated, Arabic/Persian signature recognition and retrieval. Signature region is detected using geometric properties of connected components. The description of spatial distribution of pixels in the interested region is found using angular radial partitioning scheme. This method is found better than an approach compare to the line segment distribution method. Lu and Tan [13] presented a method for information retrieval from digital libraries. Word coding techniques are employed for document image retrieval. However this method does not support integration of linguistical knowledge.

In 2005, Liu et. al [14] used density distribution features and key block features in their proposed system for document image retrieval. Key block features are used for to improve retrieval performance. This method supported retrieval of color and gray-scale document images from huge document image database of different languages. Nakai et. al [15] presented a technique for document image retrieval that is characterized by indexing with geometric invariants and voting with hash tables. This method works with high accuracy when normal digital camera is used for capturing document images. In 2006, Balasubramanian et. al [16] proposed a method to retrieve imaged documents at larger scale. Effective retrieval is achieved matching image features at word level for printed documents. Dynamic Time Wrapping (DTW) based features are used for representation of the words. Srihari et. al [17] proposed a method for signature based document image retrieval. They used signature as query for retrieval of documents. They segmented signature by removing noise and printed text from document containing signature. Global shape based binary feature vectors with normalized correlation similarity measures are employed for signature matching.

In 2007, Schomaker [20] proposed a method for retrieval of handwritten lines of text from historical documents. This method employed a brute-force matching of line-strip images for retrieving the handwritten text. Zhu and Doermann [19] presented an automatic document logo detection and extraction from document images. Boosting strategy across multiple image scale is used for classifying and localizing the logo. Spitz described character shape codes for detection of duplicate document [5], retrieval of information [6], word recognition [22], and reconstruction of document [18] without resorting the character recognition. Character cells are first segmented to obtain character shape code. This method fails with words containing connected characters. Joutel et .al [21] proposed a technique for classification of ancient manuscripts using curvelets based feature extraction. This technique achieved language independent, visual orientation and appearance based document image retrieval.

In 2008, Lu et. al [22] used word shape coding for document image retrieval. They retrieved document image by a new word shape coding scheme that captures the document content through annotating each word image by a word shape code. The word image is annotated by a set of topological shape features with character ascenders/descenders, character holes and character water reservoirs. The method proposed is found fast, efficient and tolerant to various types of document degradation. Zhu et. al [23] presented a signature-based document image retrieval. The system is able to detect and segment signature automatically during retrieval of document images and it works with unconstrained layout and complex backgrounds.

In 2009, Wang and Chen [26] used boundary extension of feature rectangle in their proposed method for logo detection. This method works on assumption that logos have white background and is independent of shape of logos. They used a decision tree classifier to detect a logo from possible logo candidates. This reduced the false positive from the logo candidate pool. Zhu and Doermann [25] proposed an automatic logo-based document image retrieval system. Logo detection and segmentation is done by using a cascade of classifiers across multiple image scales. For logo matching they employed translation, scale and rotation invariants of shape descriptors. This approach is segmentation free and layout independent. Hassan et. al [27] presented document image indexing and symbol recognition using shape descriptors. Hierarchical distance based hashing technique is used in this approach for document image indexing. Li et. al [28] proposed a system for document image retrieval with local feature sequences. The image retrieval method used is fast and OCR-free. They used local features and intrinsic, unique, page-layout free characteristics of document images. In 2010, Kokare and Shirdhonkar [30] presented comprehensive survey on document image retrieval system that provides an insight of state of art of research activities from 1992 to 2009. The paper also describes the methods, applications and challenges involved in document image retrieval.

In 2011, Hassanzadeh and Pourghassem [32] employed spatial and structural features in their proposed method for logo detection and recognition. It considers some specifications of logo such as centroid coordinates and intersection of each logo's separated part in detection process. Separated parts of logos are combined using dilation operation. A new feature based on histogram of object occurrence in a logo image is used for recognition purpose. Shirdhonkar and Kokare [33] proposed a method for document image retrieval using signature as a query. Rotated complex wavelet filters (RCWF) and dual tree complex wavelet transform (DT-CWT) are used to extract features of signature. Canberra distance and relevance with query point movement (QPM) is used during retrieval of documents to improve performance. Shirdhonkar and Kokare [34] also presented a technique for handwritten document image retrieval system. The distance measures Canberra distance and Euclidean distances are used for similarity measure in the proposed system. Superiority of Canberra distance is observed over Euclidean distance in terms of average retrieval rate.

In 2012, Keyvanpour and Tavoli [35] proposed feature weighted technique for improving performance of document image retrieval system. Feature weighting is an approach that approximates the optimal degree of influence of individual features of document images. This method weights the feature using coefficient of multiple correlations.

In 2013, Pirlo et. al [38] presented a method for layout based document image retrieval using dynamic time warping for commercially designed documents. Morphological operations are used for extracting grid based structural components from the document image. Random transform is used for description of the layouts and dynamic time warping is used for document indexing. Shekhar and Jawahar [39] proposed document specific sparse coding for word retrieval. They compared the visual similarity between two images using Bag of words (BoW). Performance of the method is improved by defining a document specific sparse coding scheme for representing visual words (interest points) in document images. This method is motivated by successful use of sparsity in signal presentation by exploiting the neighborhood properties. Keyvanpour and Tavoli [37] presented document image retrieval algorithms, analysis and future directions for the researchers in the area of document image retrieval. This paper proposes a framework for classifying document image retrieval approaches and then evaluates these approaches based on important measures. A comparative study of methods for segmentation of handwritten digits is proposed by Ribas et. al [40]. An adaptive water flow model for the binarization of degraded document images is proposed by Vaizadeh and Kabir [41]. Multilayer perceptron is used to label every blob either as text or non-text in document images. This method preserves stroke connectivity.

In 2014, Cote and Albu [42] proposed a method to classify document image into four fundamental classes (text, image, graphics, and background). The method used support vector machine for classification. Serrano et. al [43] proposed a technique to implement interactive handwriting recognition, which reduces user effort in the transcription of handwritten text in (old) documents. Sankar et. al [44] worked on large scale document image retrieval by automatic word annotation. They proposed a framework, which replaces naive classification with a mixture of indexing and classification schemes. Hoang et. al [45] presented a framework for edge noise removal from bi-level graphical document images. Rusinol et. al [46] presented Multimodal page classification in administrative document image streams for banking application. Proposed architecture represents classification of administrative document images by merging visual and textual descriptions. Hierarchical representation of the pixel intensity is used for visual description and latent semantic analysis is employed for textual description. Several off-the-shelf classifiers and many other strategies are evaluated for combining visual and textual cues. A final step uses an n -gram model for finer-grained classification of pages.

4. CLASSIFICATION OF DIRS

Figure 3 shows the framework proposed in [36] for classification of non-textual document image retrieval system (DIRS). The framework classifies document image retrieval system (DIRS) into two main categories: Traditional indexing approach and new methods of indexing.

i. **Traditional indexing:** This method of indexing includes keyword spotting and document image indexing with Optical Character Recognition (OCR). These methods are used for title based searching and text to speech conversion applications.

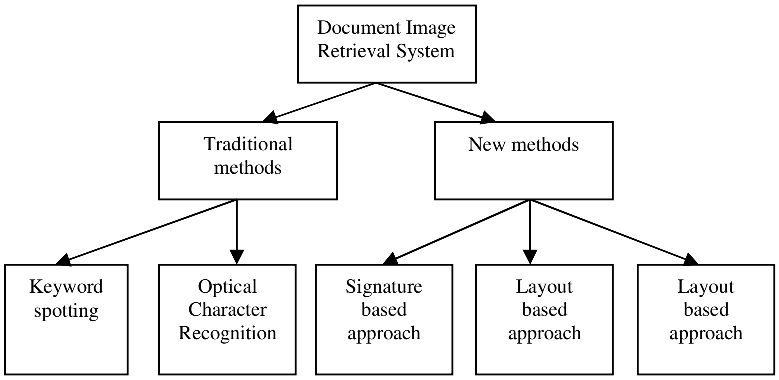

Figure 3. Classification of DIRS

ii. **New method of indexing:** This is the recent approach of indexing the documents and is based on signature, layout and logos. The layout based retrieval could be again depending on logical, physical or functional structures. Physical structures are colors, fonts, block types etc. Functional structured layout addresses indexing based function of the layouts such as address/ attachment in mails, abstract/introduction/ conclusion in scholar papers etc.

5. APPLICATIONS

The document image analysis and retrieval has lot of applications. Some of the important applications are:

5.1 Applications of Document Image Analysis

The ultimate goal of document image analysis is to develop paperless office. Document image analysis can be used in many applications. Some of the applications include:

- Reading books and documents for visually impaired by converting the text in document image to speech.
- Signature verification in banks, offices and in crime department for detecting forgery.
- Reading and recognizing number plates of vehicles using computers in vehicle registration, tracking vehicles to detect over speed and breaking of traffic rules.
- Sorting of large document datasets such as legal, historical or security related documents.
- Document image analysis can be used for designing better search engines on the Web.
- Name, Address, Location and Pin code extraction from the mails in couriers and postal department.
- Document image analysis can also be used for compression imaged documents.
- Identification of scripts in different languages is also possible with the help of document image analysis.

5.2 Applications of Document Image Retrieval

- **Title based searching:** This application helps the users to locate a specified word in document images.
- **Document similarity measurement:** It allows the user to retrieve documents by specifying entire document image as a query rather than a keyword.
- **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.
- **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.

6. CHALLENGES IN DIARS

To design and implement successful document image analysis and retrieval system following are the challenges to be addressed.

- **Computational speed:** Document image analysis and retrieval requires a sequence of steps. These steps may include document capturing, feature extraction, feature analysis, matching the features, ranking of documents. However 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.

- **Degradation of documents:** The degradation of printed or scanned document images can be due to several reasons. Some of the reasons are listed below.

    - ✓ 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.
- **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 researchers to develop and implement language independent document analysis and retrieval algorithms.
- **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.

7. EVALUATION STRATEGIES

This section describes the evaluation strategies used for document image analysis and retrieval. Document image analysis such as analysis, recognition and extraction of characters, titles, signature, logo, layout, address, labels etc. from complex documents most of the researchers consider accuracy and time required for extraction as the performance metrics.

Three important strategies Precision, Recall and Retrieval time (Speed) are used [24] as evaluation strategies for document image retrieval.

- **Precision:** Precision can be defined as the percentage of retrieved document images that are relevant to the query. This parameter can be calculated using equation (1) shown below.

  $$Precision = R_n / N \qquad (1)$$

  Where "N" is number of document images retrieved and "R_n" is number of relevant matches among retrievals.
- **Recall:** It is the ratio of relevant document images retrieved to the total number of relevant documents available in database of document images. Equation (2) can be used to calculate the recall.

  $$Recall = R_n / M \qquad (2)$$

  In the above equation "R_n" is number of relevant documents retrieved and "M" refers to the total no. of relevant documents available in the database.

- **Retrieval time (speed):** It can be defined as time required for retrieving query image by the algorithm for a finite set of database. It can be expressed in seconds or milliseconds. The retrieval time could depend on nature and size of database used for testing retrieval algorithms

8. CONCLUSION

This paper presents the technical achievements in the field of document image analysis and retrieval. It describes general frame work for document image analysis and retrieval system. This paper also highlights the applications, issues, challenges and scope for research in the domain.

REFERENCES

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

[2] F. Leclerc and R. Plamondon, "Automatic signature verification: The state of the art—1989–1993," *International Journal of Pattern Recognition, Artificial Intelligence (IJPRAI)*, vol. 8, no. 3, pp. 643–660, June 1994.

[3] 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.

[4] D.Niyogi and S. Srihari, "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, 1997.

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

[6] A. F. Smeaton and A. I. Spitz, "Using Character Shape Coding for Information Retrieval," *In Proc. Fourth International Conf. Document Analysis and Recognition*, pp.974 –978, 1997.

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

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

[9] 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.

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

[11] Rangachar Kasturi, Lawrence O'Gorman and Venu govindraju, "Document image analysis : A primer," *Sadhana,* Vol. 27, Part 1, pp. 3–22, 2002.

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

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

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

[15] Tanohiro Nakai, Koichi Kise, Masakazu Iwamura, "Camera Based Document Image Retrieval as Voting for Partial Signatures of Projective Invariants," *Proc. of the 2005 Eighth International Conf. on Document Analysis and Recognition., 2005.*

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

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

[18] A. L. Spitz, "Progress in Document Reconstruction," *In Proc. 16th Int'l Conf. Pattern Recognition*, vol. 1, pp. 464-467, 2006.

[19] Zhu, G. and D. Doermann, "Automatic document logo detection," *In conference on document analysis and recognition*, pp: 864-868, 2007.

[20] L.R.B. Schomaker, "Retrieval of Handwritten Lines in Historical Documents," *Document Analysis and Recognition, ICDAR 2007*, vol. 2, pp. 594-598, 2007.

[21] Guillaume Joutel, Veronique Eglin, Stephane Bres and Hubert Emptoz, "Curvelets based features extraction of handwritten shapes for ancient manuscripts classification," *Proc. of SPIE-IS&T Electronic Imaging*, SPIE vol.6500, 2007.

[22] Shijian Lu, Linlin Li and Chew Lim Tan, "Document Image Retrieval through Word Shape Coding," *IEEE Trans. Patt. And Mach. Intell.*, vol. 30, no.11, pp.1913-1918, 2008.

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

[24] Ritendra Datta, Dhiraj Joshi, Jiali, and James Z. Wang, "Image Retrieval: Ideas, influences, and Trends", 2008.

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

[26] Hongye Wang and Youbin Chen, "Logo Detection in Document Images Based on Boundary Extension of Feature Rectangles," *10th International Conference on Document Analysis and Recognition*, 2009.

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

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

[29] H. Srinivasan and S. Srihari, "Signature-Based Retrieval of Scanned Documents Using Conditional Random Fields," *Computational Methods for Counterterrorism*, ISBN 978-3-642-01140-5, Springer-Verlag, Berlin, Heidelberg, pp. 17-32, 2009.

[30] Manesh B. Kokare, M. S. Shirdhonkar, "Document Image retrieval: An Overview," *International Journal of Computer Applications,* vol. 1, no. 7, pp. 128-133, 2010.

[31] Z. Li, M. Schulte-Austum and M. Neschen, "Fast Logo Detection and Recognition in Document Images," *IEEE International Conference on Pattern Recognition*, pp. 2716-2719, 2010.

[32] Sina Hassanzadeh and Hossein Pourghassem, "A Novel Logo Detection and Recognition Framework for Separated Part Logos in Document Images", *Australian Journal of Basic and Applied Sciences*, 5(9): 936-946, 2011.

[33] M.S.Shirdhonkar and Manesh B. Kokare, "Document image retrieval using signature as query," *in proc. of second international conference on computer and communication technology*, pp. 66-70, MNNIT, Allahabad, India, 2011.

[34] M.S.Shirdhonkar and Manesh B. Kokare, "Writer Based Handwritten Document Image Retrieval Using Contourlet Transform," *Advances in Digital Image Processing and Information Technology Communications in Computer and Information Science,* vol. 205, pp 108-117, 2011.

[35] M. Keyvanpour and R. Tavoli, "Feature Weighting for Improving Document Image Retrieval System Performance," *International Journal of Computer science*, vol. 9, pp.125-130, 2012.

[36] Reza Tavoli, "Classification and Evaluation of Document Image Retrieval System," *WSEAS transaction on computers*, issue 10, vol. 11, pp. 329–338, 2012.

[37] Mohammadreza Keyvanpour1 and Reza Tavoli, "Document Image Rettrieval: Algorithms, Analysis and Promising directions," *International Journal of Software Engineering and its Applications*, vol. 7, No. 1, 2013.

[38] Giuseppe Pirlo, Michela Chimienti, Michele Dassisti, Donato Impedovo, Angelo Galiano, "Layout Based Document Retrieval System by Radon Transform Using Dynamic Time Warping," *Image Analysis and Processing –ICIAP 2013*, Lecture Notes in Computer Science vol. 8156, pp 61-70, 2013.

[39] Ravi shekhar and C.V.Jawahar, "Document Specific Sparse Coding for Word Retrieval", *Document Analysis and Recognition(ICDAR), 12th International Conference*, pp. 643-647, 2013.

[40] F. C. Ribas, L. S. Oliveira, A. S. Britto Jr., R. Sabourin, "Handwritten digit segmentation: a comparative study," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 16, Issue 2, pp. 127-137, 2013.

[41] Morteza Valizadeh, Ehsanollah Kabir , "An adaptive water flow model for binarization of degraded document images," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 16, Issue 2, pp. 165-176, 2013.

[42] Melissa Cote, Alexandra Branzan Albu, “Texture sparseness for pixel classification of business document images,” *International Journal on Document Analysis and Recognition (IJDAR),* 2014.

[43] Nicolás Serrano, Adrià Giménez, Jorge Civera, Alberto Sanchis, Alfons Juan, “Interactive handwriting recognition with limited user effort,” *International Journal on Document Analysis and Recognition (IJDAR)* , vol. 17, Issue 1, pp. 47-59, 2014.

[44] K. Pramod Sankar, R. Manmatha, C. V. Jawahar, “Large scale document image retrieval by automatic word annotation,” *International Journal on Document Analysis and Recognition (IJDAR),* vol. 17, Issue 1, pp. 1-17, 2014.

[45] Thai V. Hoang, Elisa H. Barney Smith, Salvatore Tabbone , “Sparsity-based edge noise removal from bilevel graphical document images,**”** *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, Issue 2, pp. 161-179, 2014.

[46] Marçal Rusiñol, Volkmar Frinken, Dimosthenis Karatzas, Andrew D. Bagdanov, Josep Lladós, “Multimodal page classification in administrative document image streams,” *International Journal on Document Analysis and Recognition (IJDAR)*, 2014.

AUTHORS

Dr. M.S.Shirdhonkar has received B.E (CSE), M.E (CSE) from Shivaji University, Kolhapur and his Doctorate from Swami Ramanand Teerth, Marathwada University, Nanded. He has 16 years of experience. His area of interest is Image processing, Document image retrieval and analysis and Pattern recognition. He has published more than 19 papers in national and international conferences and journals.

Mr. Umesh D. Dixit has received B.E (E&C) and M.Tech (CSE) from Visvesvaraya Technological University, Belagavi. He has an experience of more than 12 years and his area of interest are image processing, Dcoument image retrieval and Embedded systems. Currently pursuing Ph.D under Visvesvaraya Technological University. Belagavi