Classification and Evaluation of Document Image Retrieval System

REZA TAVOLI[1]
[1] Department of Mathematics,
Islamic Azad University, Chalous Branch (IAUC)
17 Shahrivar Ave., P.O. Box 46615-397, Chalous, Iran
r.tavoli@gmail.com

Abstract: - Document Images are documents that normally begin on paper and are then via electronics scanned. These documents have rich internal structure and might only be available in image form. Supplementally, they may have been created by a union of printing technologies (or by handwriting); and include diagrams, tables, graphics and other non-textual component. Large collections of such complex documents are commonly found in legal investigation. Many approaches come in for indexing and retrieval document images. In this paper we proposed a framework for classify non-textual document image retrieval approaches, and then we evaluated these approaches based on important measures.

Key-Words: Document image, retrieval, indexing, information system, query image, machine-print handwriting.

1 Introduction

Digitization supplies an effective way to process, preserve, and transfer all types of information. On the other hand the question arises how to find the relevant information in a large lot of data [2], [13]. Document image retrieval is a very interesting area of research with the successive growth interesting and expanding security requirements for the evolution of the modern society [12]. Respecting the growing size of data to be searched, precision is no more the only criterion for efficiency. Speed in search plays a significant role too. In accordance with tradition document image processing system is controlled by a method called Optical Character Recognition (OCR), which has obtained very good outcomes on text reading in documents. However, beside text information some documents also include information in graphical symbols such as logo [6], signature [8], machine-print, noise, etc.

If the target is to regain relevant documents from a document image database and the exact words are not required, therefore executing OCR on the entire document body is excessively expensive. Thus, a keyword-spotting technique is proposed to make possible an end user to search the images for words and filter out the relevant documents. As one of the most penetrating graphical components in government and commerce documents, logos may make possible direct identification of organizational things and serve widely as a declaration of a document's origin and ownership [14], [16]. Suppose a great collection of documents, seeking for a special logo is a highly efficient way of retrieving documents from the collaborated organization. Constructing an efficient access to these document images needs planning a mechanism for efficient search and retrieval of data image from document image collection [6], [13]. In searching a great repository of document images, an interesting work is that of retrieving documents from a dataset use of signature as a query [10].

The rest of the paper is organized as follows. In section 2, we describe document image retrieval system architecture. We describe evaluation metrics in section 3. In section 4, the proposed framework for classify document image retrieval approaches is presented. Then in section 5, we evaluate document image retrieval approaches. And, section 6 includes the conclusions.

2 Document Image Retrieval

In Figure 1, block diagram describes the stages included in document image retrieval System [12]. The various steps included in document image retrieval are feature extraction, noise removal, and matching algorithm, which are talked about here.

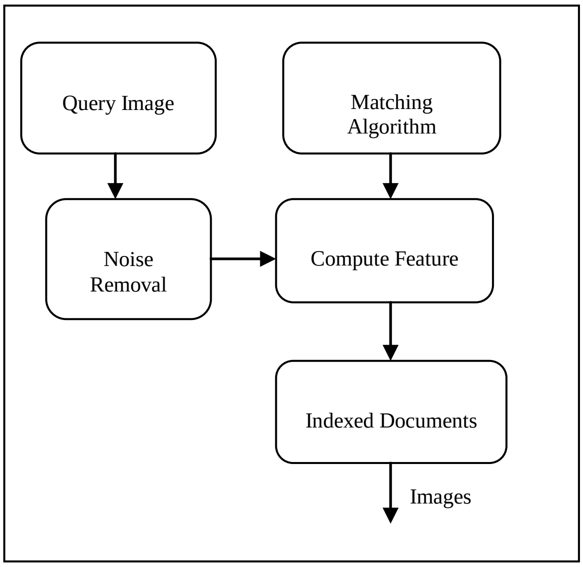

Fig 1. A Block diagram describing the steps included in document image retrieval system [12].

2.1 Query Image

Query Image is a request form end user for retrieval indexed documents**.** First end users enter query image, then system retrieval document images relevant with query image.

2.2 Noise Removal

Noise removal is fulfilled to get remove of each noise or printed text extending over the extracted images such as logos, signature, machine-print, etc. In the preprocessing stage the printed text is dismissed from the image instances. To dismiss the printed text from images various of methods can be used as an example of image improvement methods based on Support Vector Machine (SVM), chain code, to classify each connected component as a part of signature [10], noise elements, logo[6], [13], handwritten text[17], noise, etc.

2.3 Feature Extraction

Feature extraction includes extracting the significant knowledge from the document images. One time the features are extracted, they are saved in the database. One of the biggest benefits of feature extraction is that it meaningfully decreases the information to represent an image for comprehending the content of that image. It uses variety technical skills to extract the features like for example Structural, Concavity features and Gradient, which criterions the image attributes at local, medium and large scale, features based on key block features and density distribution, Angular radial partitioning of a images regions, fisher classifier, DTW, Conditional Random Field[10], etc. are used for feature extraction[12].

2.4 Matching Algorithm

The document image retrieval is executed use of similarity method to compare the query with image database [12]. Figure 1, Shows the several operative steps in the retrieval process: 1) Noise removal from the query; 2) Feature extraction from the query image; 3) Matching the query image features to each of the documents that indexed in the database 4) Sorting the documents in accord with the results from the matching method. The work of matching algorithm is to contrast the feature with the features (indexed in the database) of the document images. Similarity measure, the database feature vector and query feature vector is compared use of distance measure. The images are sorted based on the distance value. The similarity of different metrics like for example Chebychev, Euclidean, Manhattan, etc is done in. The normalized similarity is believed to be good for feature vectors as characterization to other measures.

The Euclidean distance between the features of the query image and the indexed features in the databases involved in the working document is computed with Eq1 [7]:

$$Dis(p,r)=\sqrt{\sum_{i=1}^{n}(Q(p_i)-D(p_i,r))^2} \qquad (1)$$

Where p is the feature that is being compared, D is the feature of the document (Indexed in the database); Q is the feature of the query, n is the count of component of the feature vector and r is the quantity of the document compared query. Eventually, there is a set Dis (p, r) which comprise of the Euclidean distances between each Indexed document and the query for any features which have been discussed above.

2.5 Indexed Documents

Indexed documents are documents that display to user as results.

3 Evaluation metrics

Document image retrieval is subset of information retrieval system.

Two most common and fundamental metrics for information retrieval impressiveness is precision and recall [1].

Precision (P) is the count of retrieved documents that are relevant [1]:

$$Precision=\frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})}=P(\text{relevant}|\text{retrieved}) \qquad (2)$$

Recall (R) is the count of relevant documents that are retrieved [1]:

$$Recall = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant Items})} = \mathrm{P}(\text{retrieved}|\text{relevant}) \qquad (3)$$

4 Proposed Framework For Classify Non-Textual Document Image Retrieval System

In this section we proposed a framework for classify document image Indexing approach. According to our study on document image indexing and retrieval, here, we have classified into two parts: Traditional indexing and Today Indexing. Our classification of document image indexing approaches is shown in Figure 2.

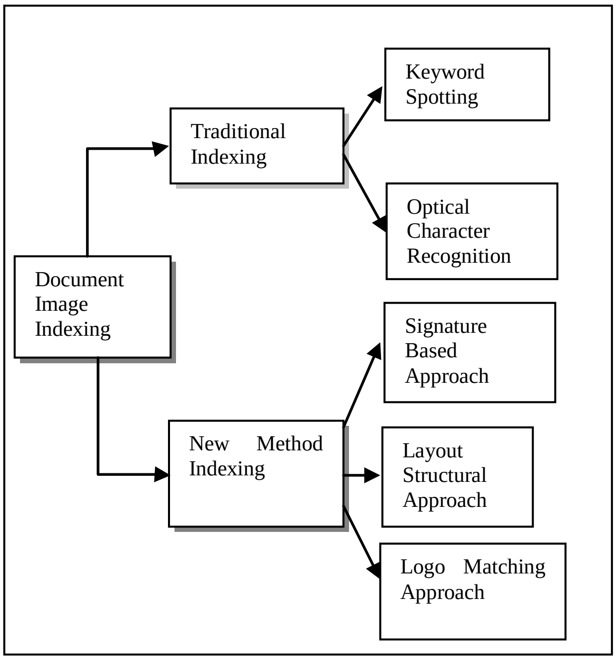

Fig 2. Proposed Framework for Classify document image indexing approaches.

Many document image indexing approaches have been proposed to improve the document image retrieval performance. Our study on Document image indexing approaches shows that document image indexing approaches can classified in to five main classes: document image indexing with OCR, document image indexing based on keyword spotting, document image indexing based on signature image, document image indexing based on layout structural and document image indexing based on Logo matching.

4.1 Document Image Indexing with Optical Character Recognition (OCR)

OCR is the electronic or mechanical translation of scanned images of handwritten, printed (numerals, letters, and symbols) into computer-processable format. OCR makes it could be to store the text, search for a word or phrase and edit the text more in compact form, show a copy free of scanning objects, and implement methods such as text mining, machine translation and text-to-speech to it [19], [22]. Figure 3, depicts the overall structure of the Document Image Retrieval System Base on OCR.

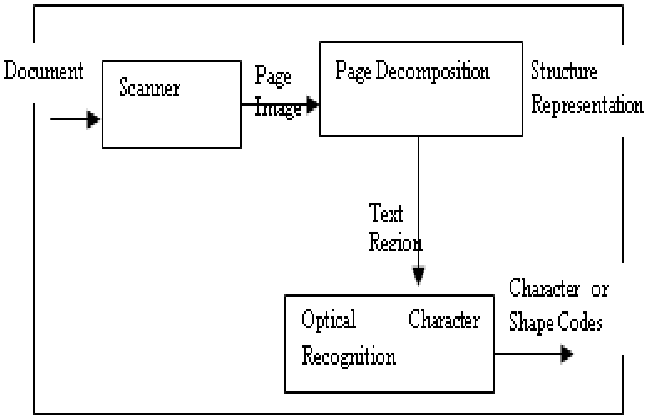

Fig.3 The overall structure of the Document Image Retrieval System Base on OCR [26].

In [20], Steven et al supply a short survey of work done to enhance the efficiency of retrieval of OCR text. Their general equation for the Retrieval Status Value (RSV) in OCR of a document is given in Equation 4 below:

$$RSV(q,d_j) = \sum ff(\varepsilon_i,q)\, ff(\varepsilon_i,d_j)/\lambda_j \qquad \textbf{(4)}$$

Where dj is document, q is query, ff (ε_i, d_j) is feature frequency in the document, ff (ε_i, q) is feature frequency in the query, λ_i is count of happening of feature frequency in document. In 1996, the meeting kept a disorder track where the test data was achieved by printing, scanning and recognizing via OCR information from the Federal Registry [5].

4.2 Document Image Retrieval Based On Keyword Spotting

Figure 4, shows the comprehensive diagram of the DIRS Base on Word Spotting [7]. It is composed of two sections: The online and the offline procedure. In the offline procedure the repository of document images are tested and the results are saved in a database. This digital scanning composes of 3 steps. At initial step the document transports the preprocessing step which involves a binarization with the skeletonization, a mean filter and Otsu method. The word segmentation step is subsequent the preprocessing step. Its fundamental target is to discover the word blocks. In the last step of the offline procedure the features of any word are computed and saved in the database [7].

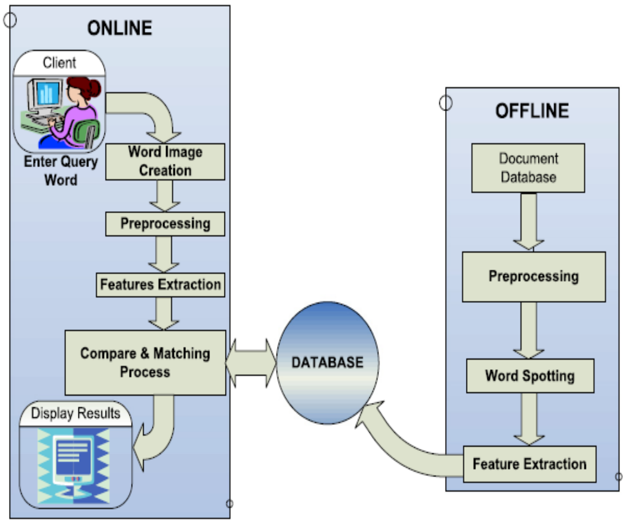

Fig.4 Comprehensive diagram of the DIRS Base on Word Spotting [7]

The online procedure composes of the interface for end user can control the system (input the query, view the outcomes), the construction of the word's image, preprocessing and features extraction steps which are the same thing as mentioned before, and at last, the similarity process of the query features with indexed features in the database. For each word image use Width to Height Ratio, End Points, Cross Points, etc as Features.

Many interesting work has been done on the problem of looking for keywords in document images use of alone image characteristic [5], [9], [14] and [16]. In [14], based on the type and position of the features, a succession of feature vectors is explained to any word. The x-centroid of each black or white run block is initially computed by the subsequent Equation 5:

$$C_x = \frac{\sum_{i=1}^{n}(z_i - y_i)\times x_i}{\sum_{i=1}^{n}(z_i - y_i)} \qquad (5)$$

Where C_x is x-centroid of each white or black run block, *(x_1, y_1, z_1)* to *(x_n, y_n, z_n)* equals to the farthest to the left and farthest to the right run in the white run block.

4.3 Document Image Retrieval Based on Layout Structural Similarity

For great databases, physical indexing (or manual indexing) can be preventively costly, not to indicate the personal and perhaps nearsighted explanation by the person constructing the index, and the restricted significance of keywords. As a result, the problem of automatic indexing of document images by content has developed as an interesting field of research [11]. Arrangement of document images separate to two kinds: conceptual and Geometrical [11]. Figure 5 also depicts three types of features:

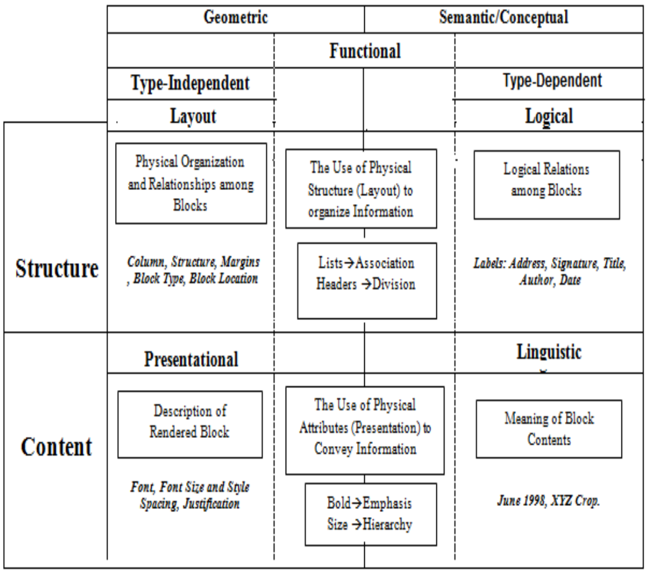
Geometric
Semantic/Conceptual
Functional
Type-Independent
Type-Dependent
Layout
Logical
Structure
Physical Organization and Relationships among Blocks
The Use of Physical Structure (Layout) to organize Information
Logical Relations among Blocks
Column, Structure, Margins , Block Type, Block Location
Lists→Association Headers →Division
Labels: Address, Signature, Title, Author, Date
Content
Presentational
Linguistic
Description of Rendered Block
The Use of Physical Attributes (Presentation) to Convey Information
Meaning of Block Contents
Font, Font Size and Style Spacing, Justification
Bold→Emphasis Size →Hierarchy
June 1998, XYZ Crop.

Fig 5.The geometric, semantic content and structure descriptions [11].

Logical Structural: The remarked types in the above examples, letter and memo are logical types, and their member objects are logical objects. Logical structure is subset of conceptual type.

Physical Structural: Physical structure composed of natural features for example annotation, Color, Font, Block types, etc. Physical structure is subset of geometrical type.

Functional Structural: Functional structure demonstrates entirely in Table 1. Functional structure is between of geometrical type and logical type.

Table 1: Functional Structure

Structure	Example	Use
Header	Centered	focal point ,Relative importance
List	Enumerated, itemized	Conveys temporal sequence Suggest similar level of descriptiveness
Separator	White space or rule line	Physical and possibility semantic disassociations
Attachment	Boxed text Side bar Footnote	Supplemental information under some semantic hierarchy
Illustration	table Figure	Supplemental information-Preserves 2D association. Graphics representation of info

Structural layout analysis can be executed in bottom-up or top-down mode [22].

4.4 Signature Based Document Image Retrieval

In searching compound documents, like for example achieves of office documents, a work of pertinence is narrating the signature in a specific document to the nearest similar within a database of documents; this is famous as signature retrieval role. Suppose a database of document (signed document), it will be interesting to narrate an asked document to another documents in the database that have been signed by the previous author [12]. Figure 6 depicts Indexing documents with signature.

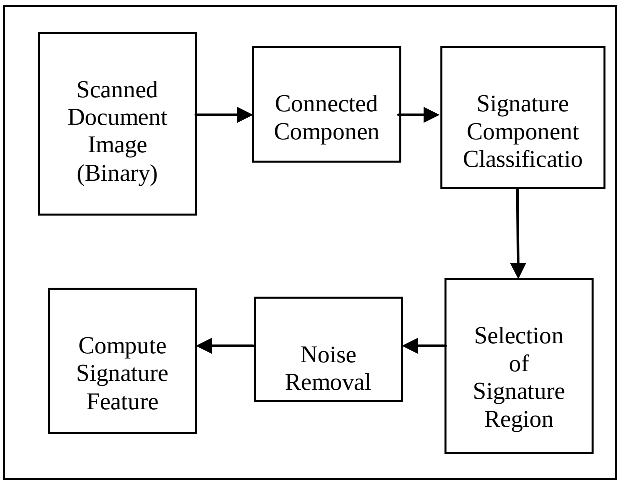
Scanned Document Image (Binary)
Connected Componen
Signature Component Classificatio
Selection of Signature Region
Noise Removal
Compute Signature Feature

Fig 6. Indexing Documents with Signature [10]

Document image retrieval base on Signature composed of 3 stages [10]: Stage1: Extraction the block of signature; Stage1-1: Extraction the Connected Component; Stage1-2: Classification the Signature Components; Stage1-3: Signature Region Selection; Stage2: Removal the Noise in the block of the signature; Stage3: Extraction the feature vector of Signature.
Figure 7 depicts Stage1 to Stage3:

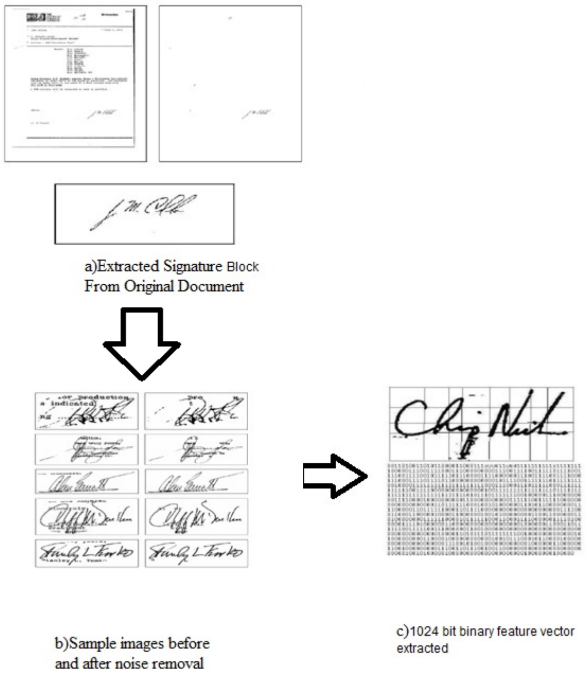

Fig 7. Sample Indexing document with signature results: a) Step1:Signature Block Extraction; b) Step2:Noise Removal c)Step3:Signature Feature Extraction[10]

Retrieval: Figure 8 depicts the variety operational stages in the retrieval procedure: (a) noise removal from the query signature; (b) feature extraction from the query signature; (c) similar the feature vectors of query signature to each of the feature vectors indexed in the database; and (d) sorting the documents in accordance with the results from the matching algorithm [10].

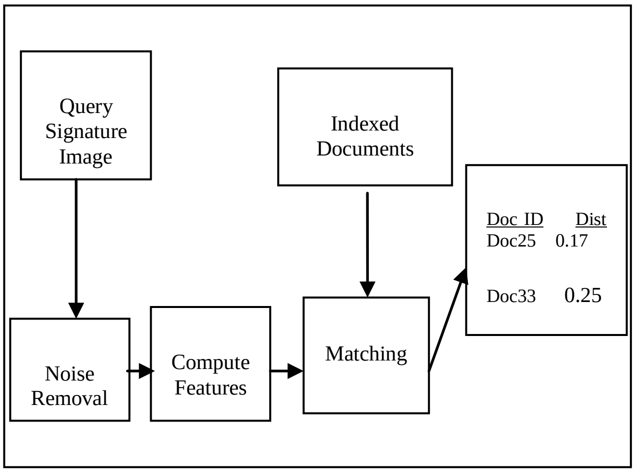

Fig 8. Operational steps in the retrieval process [10]

Matching algorithm: Figure 9 depicts a query signature image being matched versus small number of signatures and the resulting dissimilarity metrics achieved use of the similarity method. The distance between the queried signature and any documents (indexed in the database) is computed use of normalized correlation similarity metric. Suppose two feature vectors P $\in \Omega$ and $Q \in \Omega$, each similarity grade S (P, Q) uses all or some of the four suitable estimate, i.e. S_{00}; S_{01}; S_{10}; S_{11}. The similarity distance S(P, Q) between two feature vectors P and Q is given by Eq6 [10].

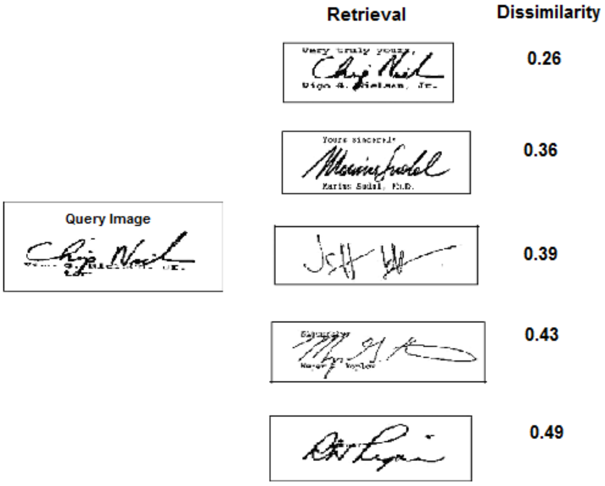

Fig 9. Some of results with the query on the left and the signatures matched versus and their comparable dissimilarity distances on the right [10].

$$S(p,q) = \frac{1}{2} + \frac{S_{00}S_{11} - S_{10}S_{01}}{2((S_{10}+S_{11})(S_{01}+S_{00})(S_{11}+S_{01})(S_{00}+S_{10}))^{\frac{1}{2}}} \quad (6)$$

Where
S_{00} = the first vector has a 0 and the second vector also has a 0 in the similar locations.
S_{11} = the first vector has a 1 and the second vector also has a 1 in the similar locations.
S_{01} = the first vector has a 0 until the second vector has a 1 in the similar locations.
S_{10} = the first vector has a 1 until the second vector has a 0 in the similar locations.

4.5 Document Image Retrieval Based On Logo Matching

One of the largest amount penetrating graphical components in commerce and government documents, logos may make possible direct detection of organizational things and serve widely as a proclamation of a document's ownership and origin [6]. Suppose a great collection of documents, seeking for a special logo is a highly efficient way of retrieving documents from the collaborated organization. Constructing an efficient access to these document images needs planning a mechanism for efficient search and retrieval of data image from document image collection [12].
Figure 10 depicts the overall structure of logo identification and recognition in document images. In order to identify and recognize this type of logos, the next contributions have been performed:1) A general logo identification and recognition arrangement with 3 layers; 2) A simple and suitable feature design; 3) A new geometrical recreation algorithm.

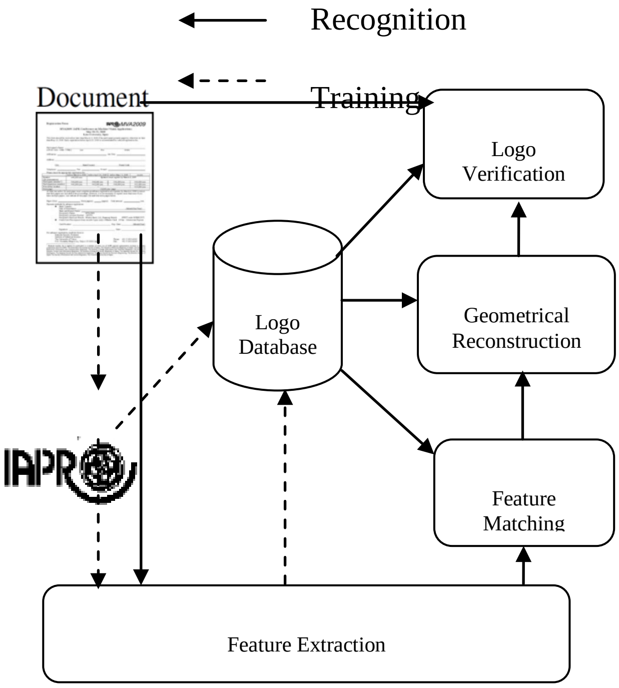

Fig 10. System architecture of logo detection and recognition [13]

In [23], [24] and [25] some approaches to logo detection and recognition have been proposed.

5 Evaluations of Document Image Retrieval Approaches

In this section, we evaluate approaches based on important measures. Our evaluation is summarized in Table 2.

The measures that considered in our evaluation of document image retrieval and indexing approaches are as follows:

Application Type: A document image retrieval approach has various applications such as similarly documents, word searching, duplicate detection, etc.

Appearance Features: A document image retrieval approach has many appearances features. Any approach has specified appearance feature for it.

Query Image: any approach has query image for retrieval documents such as signature image or word image**.** First users enter query image, then system retrieval document images relevant with query image.

Is Structural: which one is approach for document image retrieval structural? Meaning of Structural is table and formatting in document images.

Language Independent: which one is approach for document image retrieval language independent?

Cost: 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 cost expensive. Each of the approach has different cost for matching and retrieval.

Techniques: Each of the approach has various techniques for indexing and retrieval documents.

Problems: Complex documents pose a great challenge in the field of document recognition and retrieval. Each of the approach has different problems such as noisy data, uncommon fonts, etc.

Table 2. Evaluate approaches based on important measures

Approaches		Measures							
		Application	Appearance Feature	techniques	Problems	Image Query	Is Structural	Language Independent	Cost
Traditional Indexing	OCR	translation of scanned images into machine encoded text, full text search[3,5], categorization of Text[20]	ASCII Code	Convert To ASCII[5], Confusion Matrix[20]	Degradation of documents [20] , character Confusion , Uncommon Fonts	Word image	NO	NO	High
	Keyword Spotting	Summarization[5], Similarity Document[4], Information filtering[15], Filing System[15], processing handwriting documents[17] , duplicate document detection[15] , Retrieving imaged documents in digital libraries[21]	Word Shape Image[14,16], Character Stroke Features[5]	Work Shape/HMM [5], Word Image Matching[14,16], Character Stroke[5] Features, Shape Coding[5,16]	Filter proper nouns in images of text[5], Automatic Abstracting of Images[5]	Word image	NO	NO	Medium
New Method Indexing (Indexing Without Conversion)	Signature Based Approach	Indexing and Retrieval office documents[10]	Relative location[10], Convex hull distance[10], aspect ratio[10], number of Components[10]	Conditional Random Field[10], dynamic time warping (DTW)[18]	Noisy Data(handwritten text, machine-print,…)[8,10,12]	Signature Image	NO	YES	Low
	Layout Structural Approach	Structural Similarity of documents[11], Office Documents, Government Documents, Structural documents etc.	Conceptual(Logical Structure), Geometrical(Physical Structure & Functional Structure)[11]	query by example and sub-image matching using query by sketch [11]	Noisy Data	Document Structure	YES	YES	Medium
	Logo Matching Approach	Indexing and Retrieval business & government documents[13,24]	coarse scale level [23], boundary extension of feature rectangle[13,24]	Bag of words[25], anchor line[13]	Scale and Rotate Logo, Noisy Data(handwritten text, machine-print,…)[13,23,24]	Logo Image	NO	YES	Low

According to table 2 for searching specific word in document images is used usually from OCR and keyword spotting method. Although keyword spotting method is used word image characteristic such as Word Shape Image and character Stroke , it has more flexibility and has better behavior against noise. Also for searching in official documents such official letter the best method is signature based Approach and for governmental or organizational document the best is logo matching approach. Sometimes if searching is base on structure of document image, the best way is layout Structural because it divides the document image to three section such as physical, logical and Functional.

6 Conclusion

Traditionally, transmittal and storage of data have been by paper documents. In days gone by few ten years, documents more and more begin on the computer, but, despite this, it is vague whether the computers has enlarged or reduce the quantity of paper. Despite the fact that the concept of raw document image retrieval is interesting, inclusive resolutions which do not demand finish and exact conversion to a machine-readable form continue to be evasive for feasible systems. Many approaches come in for indexing and retrieval document images. In this paper we proposed a framework for classify document image retrieval approaches, and then we evaluated these approaches based on important measures. Our study on Document image indexing approaches shows that non-textual document image indexing approaches can classify in to three main classes: document image indexing based on signature image, document image indexing based on layout structural and document image indexing based on LOGO matching.

References:

[1]Christopher D. Manning, Prabhakar Raghavan, Hinrich Schultz. *An Introduction to Information Retrieval*, Cambridge University Press Cambridge, England, 2009.

[2]Omid E.Kia. *Document Image Compression and Analysis*, Submitted of the faculty of the Graduate school of the University of Maryland at college park in partial fulfillment of the requirements of the degree of Doctor of Philosophy, 1997.

[3]C. Barges. A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2, 1998, PP: 121-167.

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

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

[6]Guangyu Zhu and David Doermann. Logo Matching for Document Image Retrieval, *10th International Conference on document Analysis and Recognition*, 2009, PP: 606-610.

[7]K. Zagoris, N. Papamarkos and C. Chamzas. Web Document Image Retrieval System Based On Word Spotting, *IEEE International Conference on Image Processing*, 2006, PP: 477-480.

[8]Guangyu Zhu, Yefeng Zheng, and David Doermann. Signature-Based Document Image Retrieval. *ECCV*, 3, LNCS 5304, 2008, PP: 752-765.

[9]Million Meshesha · C. V. Jawahar. Matching word images for content-based retrieval from printed document images, *International Journal on Document Analysis and Recognition*, 11, 1, 2008, PP: 29-38.

[10]Harish Srinivasan and Sargur Srihari. Signature-Based Retrieval of Scanned Documents Using Conditional Random Fields, Computational Methods for Counterterrorism, *Volume. ISBN 978-3-642-01140-5. Springer-Verlag Berlin Heidelberg*, 2009, PP: 17-32.

[11]Christian Shin, David S. Doermann. Document Image Retrieval Based on Layout Structural Similarity, *IPCV*, 2006, PP: 606-612.

[12]Manesh B. Kokare, M.S.Shirdhonkar. Document Image Retrieval: An Overview, I*nternational Journal of Computer Applications* (0975 – 8887) 1, 7, 2010, PP: 114-119.

[13]Zhe Li, Matthias Schulte-Austum, and Martin Neschen. Fast Logo Detection and Recognition in Document Images, *International Conference on Pattern Recognition*, 2010, PP: 2716-2719.

[14]Shuyong Bai, Linlin Li and Chew Lim Tan. Keyword Spotting in Document Images through Word Shape Coding, *10th International Conference on Document Analysis and Recognition*, 2009,PP: 331-335.

[15]A. Lawrence Spitz, Duplicate Document Detection. *International Society for Optical Engineering, Document Recognition IV*, San Jose, 1997, PP: 88-94.

[16]Shijian Lu, Linlin Li, chew lim tan. Document Image Retrieval through Word Shape Coding, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30, 11, 2008, 1913-1918.

[17]Sophea PRUM, Muriel Visani Jean-Marc Ogier. On-line Handwriting word recognition using a bicharacter model, *International Conference on Pattern Recognition*, 2010, PP: 2700-2703.

[18]Zhang, B., S. N. Srihari, and C. Huang. Word image retrieval using binary features. Document Recognition and Retrieval XI, SPIE, San Jose, CA, .2004, PP: 45-53.

[19]Optical Character Recognition from Wikipedia, the free encyclopedia.

[20]Steven M. Beitzel, Eric C. Jensen, David A. Grossman. A Survey of Retrieval Strategies for OCR Text Collections, *Proceedings Symposium on Document Image Understanding Technology*, 2003.

[21]Simone Marinai.A Survey of Document Image Retrieval in Digital Libraries, *9th Colloque International Francophone sur l'Ecrit at el Document (CIFED)*, 2006,PP:193-198.

[22]Lawrence O'Gorman, Rangachar Kasturi. *Document Image Analysis*, IEEE Computer Society Executive Briefings, Book,2009.

[23]G. Zhu and D. Doermann, Automatic document logo detection, in *ICDAR '07: Proc. of Int. Conf. on Document Analysis and Recognition*, Washington, DC, USA, 2007,PP:864–868.

[24]H. Wang and Y. Chen. Logo detection in document images based on boundary extension of feature rectangles, *in ICDAR '09: Proc. of the Tenth Int. Conf. on Document Analysis and Recognition,* Barcelona, Spain, 2009,PP: 1335–1339.

[25]M. Rusinol and J. Llados. Logo spotting by a bag-of-words approach for document categorization, in ICDAR '09: Proc.of the Tenth Int. Conf. on Document Analysis and Recognition, Barcelona, Spain, 2009, PP: 111–115.

[26]David Doermann, *Document Images and EDiscovery*, lecture, 2009.