Document Image Retrieval: Algorithms, Analysis and Promising Directions

Mohammadreza Keyvanpour¹ and Reza Tavoli²

¹Department of Computer Engineering, Alzahra University, Tehran, Iran
²Department of electrical and computer, Qazvin Islamic Azad University (QIAU), Qazvin, Iran
keyvanpour@Alzahra.ac.ir, r.tavoli@gmail.com

Abstract

During the last decades, Due to the advances in Information technology and communication and increase in volume of printed documents in many applications, document image databases have become increasingly important. Document Images are documents that normally begin on paper and are then via electronics scanned that move towards a paperless office and stored documents as images. Document Image retrieval is one of an important research area in the field of document image databases. Many approaches come in for indexing and retrieval document images. Traditionally, Optical character recognition (OCR) has been used for completely convert the manuscript to an electronic version which can be indexed automatically. Then, Keyword spotting has been proposed for indexing document image retrieval. Keyword spotting method has lower cost than OCR. But there are some problems in both of methods for indexing document images with non-text components. Three approaches have been presented to solve this problem, Signature based approach, layout structural and logo based approach. In this paper we proposed a framework for classify document image retrieval approaches, and then we evaluated these approaches based on important measures.

Keywords: Document image retrieval, indexing, information retrieval, query image

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.

![Figure 1. A Block Diagram Describing the Steps Included in Document Image Retrieval System [12]](image)

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]:

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

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.


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]:

\[
Precision = \frac{\#(\text{Relevant \ Items Retrieved})}{\#(\text{Retrieved \ Items})} = P(\text{relevant | retrieved}) \tag{2}
\]
Recall (R) is the count of relevant documents that are retrieved [1]:

\[ \text{Recall} = \frac{\#(\text{Relevant Items Retrieved})}{\#(\text{Relevant Items})} = P(\text{Retrieved | Relevant}) \]  

(3)


In this section we proposed a new 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. Traditional Indexing are based on text Component and today Indexing are based on non-text component such as logos, signatures and layout structure f document images. Our classification of document image indexing approaches is shown in “Figure 2”.

![Proposed Framework for Classify Document Image Indexing Approaches](image)

**Figure 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 Retrieval with Optical Character Recognition

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.
Figure 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:

\[
RSV(q, d_j) = \sum_{i} ff(\varepsilon_i, q) ff(\varepsilon_i, d_j) / \lambda_j
\]

Where \(d_j\) is document, \(q\) is query, \(ff(\varepsilon_i, d_j)\) is feature frequency in the document, \(ff(\varepsilon_i, q)\) is feature frequency in the query, \(\lambda_i\) 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].

Figure 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]. 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:

\[ C_x = \frac{\sum_{i=1}^{n} (z_i - y_i) \times x_i}{\sum_{i=1}^{n} (z_i - y_i)} \]  

(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.

The approach of Chen et al. is segmentation- and recognition- free, and has applications in the information retrieval domain when Boolean models are used, as well as for information filtering. Chen et al use word image Properties as input to a Hidden Markov Model (HMM) [5]. DeCurtins and Chen use word shape information and a voting technique to execute matching of words. The method is based on features excluding blanks, vertical strokes, horizontal strokes, bowls and ovals extracted from a neighboring line of text [5]. Trenkle and Vogt explain a preliminary experiment on word level image matching. In their approach, a query term is extended to include its variations, and an image is generated in each of several fonts and with lower and upper-case characters [5].

In [16, 27, 28, 29] present several word image retrieval approaches based on word shape coding. In These methods, annotation method and its applications to the document image retrieval by either query keywords or a query document image are presented. In [30] we present a feature weighting method for Document Image Retrieval System (DIRS) based on keyword spotting. In [16] Shijian Lu et al a new method for annotate word images by a set of topological character shape features including character descenders/ascenders, character reservoirs water, and character holes proposed. “Figure 5” illustrated three topological character features.

(a) shape  (b) shape  
(c) shape  (d) shape

Figure 5. The Three Topological Character Shape Features in use: (a) the Sample Word Image “shape”; (b) Character Ascenders and Descenders; (c) Character Holes; (d) Character Water Reservoirs [16]

In [16] document images can then be retrieved by either query keywords or a query document image based on their content similarity. To locate the query keywords
properly, they format each word image \( W \) with a unique spelling as a word record as follows:

\[
WR = [WSC \quad (p_i \quad blx_i \quad bly_i \quad w_i \quad h_i)] \\
\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \ quad
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>
<thead>
<tr>
<th>Structure</th>
<th>Example</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Header</td>
<td>Centered</td>
<td>focal point, Relative importance</td>
</tr>
<tr>
<td>List</td>
<td>Enumerated, itemized</td>
<td>Conveys temporal sequence</td>
</tr>
<tr>
<td>Separator</td>
<td>White space or rule line</td>
<td>Suggest similar level of descriptiveness</td>
</tr>
<tr>
<td>Attachment</td>
<td>Boxed text Side bar Footnote</td>
<td>Supplemental information under some semantic hierarchy</td>
</tr>
<tr>
<td>Illustration</td>
<td>table Figure</td>
<td>Supplemental information - Preserves 2D association. Graphics representation of info</td>
</tr>
</tbody>
</table>

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 7” depicts Indexing documents with signature.

![Figure 7. Indexing Documents with Signature [10]](image)

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 8” depicts Stage1 to Stage3:

![Figure 8](image)

**Figure 8. Sample Indexing Document with Signature Results:** a) Step1: Signature Block Extraction; b) Step2: Noise Removal c) Step3: Signature Feature Extraction [10]

**Retrieval:** “Figure 9” 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].

![Figure 9](image)

**Figure 9. Operational Steps in the Retrieval Process** [10]

**Matching algorithm:** “Figure 10” 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 Eq\ 8 [10].

\[
S(p,q) = \frac{1}{2} \frac{S_{00}S_{11} - S_{01}S_{10}}{2(S_{10} + S_{11})(S_{01} + S_{10})(S_{00} + S_{11})} \tag{8}
\]

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 need planning a mechanism for efficient search and retrieval of data image from document image collection [12].

“Figure 11” 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.

![System Architecture of Logo Detection and Recognition](image)

**Figure 11. System Architecture of Logo Detection and Recognition [13]**

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

### 5. Evaluation 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

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Application</td>
</tr>
<tr>
<td>Traditional Indexing</td>
<td>OCR</td>
</tr>
<tr>
<td>New Method Indexing (Indexing Without Conversion)</td>
<td>Signature Based Approach</td>
</tr>
<tr>
<td></td>
<td>Logo Matching Approach</td>
</tr>
</tbody>
</table>

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.
5. Conclusion

Traditionally, transmittal and storage of data have been by paper documents. Documents increasingly begin on the computer, but, it is unclear 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 document image indexing approaches can classified in to five main classes: document image indexing with OCR, document image indexing based on keyword spotting, signature image, layout structural and LOGO matching.

References


Authors

Mohammadreza Keyvanpour is an Assistant Professor at Alzahra University, Tehran, Iran. He received his B.s in software engineering From Iran University of Science &Technology, Tehran, Iran. He received his M.s and PhD in software engineering from Tarbiat Modares University, Tehran, Iran. His research interests Include image retrieval and data mining.

Reza Tavoli was born in Chalous, Iran in 1986. He received his B.s in software engineering from Iran University of Science & Technology, Behshahr, Iran. He received his M.s software engineering from Islamic Azad University, science & Research Branch, Tehran, Iran. Currently, He is pursuing PhD in Software engineering at Islamic Azad University, Qazvin Branch, and Qazvin, Iran. His research interests Include document image retrieval and data mining.