ISSN: 0976-3031

Available Online at http://www.recentscientific.com

International Journal of Recent Scientific Research
Vol. 8, Issue, 1, pp. 15279-15283, January, 2017

International Journal of Recent Scientific Research

Research Article

INFORMATION RETRIEVAL FROM IMAGE DATABASES

Seema Yadav*., Saurabhkumar Jain., Priya Bhanushali and Tejinder Kaur

Department of Information Technology, K.J. Somaiya Institute of Engineering and Information Technology, University of Mumbai, Maharashtra, India

ARTICLE INFO

Article History:

Received 15th October, 2016
Received in revised form 25th November, 2016
Accepted 28th December, 2016
Published online 28th January, 2017

Key Words:

Document image analysis, Segmentation, Tesseract Engine, OpenCV, Indexing, Pre-processing.

ABSTRACT

As vast amount of digital image data is getting archived by the advanced libraries, there is a requirement for an ef cient search methodologies to make them accessible according to client's data requirement. For their retrieval, it is imperative to recognize their contents. Current technologies for optical character recognition (OCR) and document analysis do not handle such documents adequately because of the recognition errors. Due to these challenges, computer is unable to recognize the characters while reading them. In this paper, we propose and effective word image matching scheme that achieves high performance in the presence of noise in image, degradation and word form-variants. Initially, each image in image-database is pre-processed. In the next step find contour method is used to detect blobs which are further passed in tesseract engine. Tesseract segments the characters from the image and stores in character database. Each word in the database is used to index a given set of images. During retrieval, the query word presented to the system is matched with characters in the database and all images containing instances of the query word are retrieved and presented to the user. Using this approach, our method is able to successfully handle images with different font styles, size and heavily touching characters. From the experimental results on the variety of image database it is observed that the extraction of text from the images is mostly accurate and indexing of words based on the position is working perfectly.

INTRODUCTION

With capacity getting to be distinctly less expensive and imaging gadgets turning out to be progressively popular, efforts are on the way to digitize and archive large quantity of text and image. Success of text image retrieval systems mainly depends on the performance of optical character recognition (OCR), which convert scanned document images into texts [Million Meshesha and C. V. Jawahar, 2008]. Due to the noise and the poor contrast in the images, many extraction features must be acquired to distinguish text from complex document image. Secondly, it is difficult to recognize the text accurately. Word recognition is much more difficult because OCR blunders may incorporate version operations, for example, characters substitution, deletion, and insertion.

The objective of this system is to plan an IR strategy to extract extensive character databases and give back the archives that the framework considers applicable to the user's query. Specifically, we will take into account the possible recognition errors using the retrieval process.

This approach uses a method of text recognition from a database where all the scanned images are stored. This database will be used to retrieve the result based on the user query. Before displaying the final result, the scanned image will be pre-processed. Pre-processing operations like erosion, dilation, smoothing and thresholding are performed to remove noise for efficient data retrieval. A Find Contour method is used to detect blobs which are further passed in the tesseract engine. In tesseract engine, text segmentation is done and the extracted characters are stored into character database. Text Stream is generated based on textual information passed into the engine. Therefore, based on the user query the result is retrieved. Images with the query word are highlighted.

This paper presents a brief overview of Information retrieval from scanned image documents. The further sections explain about the related work done on the topic, our proposed methods for system, list of modules implemented in the project, feasibility study and applications on which system can be used.

Related Work

Fast access to data is a noteworthy progression acquired through computerized innovation where data is digitized and made accessible online to all partners. Be that as it may, there is still a huge document base in printed format in libraries and

Corresponding author: **Seema Yadav*
Department of Information Technology, K.J. Somaiya Institute of Engineering and Information Technology, University of Mumbai, Maharashtra, India

keeping in mind the end goal to make these available to all, computerized libraries assume a crucial part. The idea of a computerized library is not restricted to simple filtering of books and reports. These filtered records should be complemented by an information retrieval framework permitting readers rapid access to the queried information. Optical Character Reader (OCR) is one of the solutions which have matured significantly for many languages around the globe [Raashid Hussain, Haris Ahmad Khan, Imran Siddiqi, Khurram Khurshid, Asif Masood, 2015]. An attractive solution to this problem is the use of word spotting where queried information is searched by matching the word shapes instead of converting it into text. For example, Ho et al proposed a word recognition method based on word shape analysis without character segmentation and recognition [Tin Kam Ho, Jonathan J. Hull and Sargur N. Srihari,1992]. Every word is first partitioned into a fixed 4*10 grid. Features are extracted and their relative locations in the grid are recorded in a feature vector. The city-block distance is used to compare the feature vector of an input word and that of a word in a given lexicon. This method can only be applied to a fixed vocabulary word recognition system and has no ability of partial word matching. Several approaches have been proposed to enhance OCR accuracy and text detection. Some approaches tried to correct OCR errors after detecting them. Kukich proposed to use a dictionary or n-gram based approaches to detect OCR errors and replace them with the most likely word in the dictionary using statistical measures [Pramod Sankar K, R Manmatha and C V Jawahar, 2014]. These methodologies can lessen the general OCR error rates for the frequent words of the language, yet it is probably going to degenerate effectively recognized words which are not in the dictionary, for example names and places. As an alternative Tong et al proposed to use the context of the text itself to correct misrecognized words [X. Tong and D. A. Evans, 1996]. The success of these approaches depends on the language models and trained dictionaries and can be useless if used on different corpora with different vocabularies. A more recent alternative is to combine multiple OCR outputs to locate and fix OCR errors automatically without using language specific information. Other approaches are based on edge detection, binarization, connected-component based and texture-based methods [D. Doermann, 1998]. In, the authors demonstrate that best results were achieved using edge-based text detection compared to mathematical morphology and colour-based character extraction.

PROPOSED METHODOLOGY

The proposed system consists of six main modules. A database consisting of scanned images is stored and the result is retrieved based on the user query. Before displaying the final result the scanned image will be pre-processed. Pre-processing operations like erosion, dilation, smoothing and thresholding are performed to get image ready for efficient data retrieval. Erosion and dilation operations are used to increase and decrease the object boundaries. To clean the object boundaries smoothing is applied and to increase the contrast of image thresholding is carried out. A find contour method is used to detect blobs which are further passed in the tesseract engine. Tesseract engine is used for segmentation and extraction of text from images. Tesseract efficiently handles extraction of text from white on black images. Blobs are organized into text lines which are broken into words differently according to character spacing. Recognition of these words is a two-pass process where each word is passed to an adaptive classifier as training data. The text Stream is generated based on information passed into the engine. Extracted text from images is stored in the character database. The words are indexed based on the position of character in the database. When the query is entered by the user, it is matched with the training data present in the database. The matched query is indexed based on its location and the word is highlighted. This method is able to successfully handle the problem of heavily touching characters.

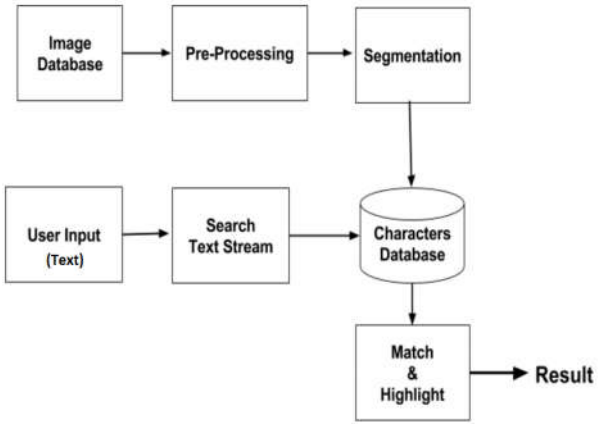


Figure 1 System Diagram for searching user specified word in document image

As image processing is one of the base domain of the system. So Open CV library is integrated in the system for processing of image documents. Open CV, is basically a library of functions written in C/C++. Programs written in Open CV run much faster than similar programs written in Matlab because machine language code is directly provided to the computer to get executed. So, conclusion is that Open CV is fast when it comes to speed of execution.

Table 1 Index Scores comparison between Open CV and MATLAB

	MATLAB	OPENCV
Ease of use	9	3
Speed	2	9
Resources Needed	4	9
Cost	4	10
Development Environment	8	6
Memory Management	9	4
Portability	3	8
Development of useful programming skills	3	8
Help and sample code	8	9
Debugging	9	5
Total	59	71

List of modules

User Interaction

The user interacts with the system by entering the query of his interest. After processing the query, the documents containing

the query word is listed, from which user can access the documents of interest.

Pre-processing

Pre-processing is a significant step where a set of all documents are gathered and passed to the word extraction phases. Pre-processing techniques are required in refining color, grey-level or binary document images containing text and/or graphics. In character recognition systems most of the applications use grey or binary images since dealing with color images is computationally high. Such images may also contain non-uniform background and watermarks making it difficult to extract the document text from the image without performing some kind of pre-processing. Therefore, the desired result from pre-processing is a binary image containing text only. To achieve this, several steps are needed. Firstly, some image enhancement techniques are used to remove noise or to correct the contrast in the image [Ankit Sharma, Dipti R Chaudhary, 2013]. Secondly, thresholding is performed to remove the background containing any scenes, watermarks and noise. Thirdly, character segmentation is done to separate characters from each other. Finally, morphological processing helps to enhance the characters in cases where thresholding is obligatory. Also other pre-processing techniques help to enhance eroded parts of the characters by adding pixels to them.

Segmentation

The segmentation is the most important process in text recognition [Mr. Pratik Madhukar Manwatkar, Mr. Shashank H. Yadav, 2015]. Segmentation is done to make the separation between the individual characters of an image and is one of the most extensive phases in this project. The performance of this project is dependent on segmentation. Images will be fed to Tesseract Engine where it performs character segmentation and these characters are stored in character database for matching of query word. Segmentation subdivides an image into its constituent regions or objects. Basically in segmentation, we try to extract basic constituents of the script, which are certainly characters. This is needed because the system classifier recognizes characters only.

Segmentation – Tesseract Engine

Tesseract is an example based system which means that the engine works on a set of example rules defined in the system and results depend on this data. To get good results, it is necessary to define these set of rules accurately which is called "Training the engine". The reason to flexibility of Tesseract is the fact that we could always change or modify the rules depending on the requirements. Tesseract is an elegant engine with various layers. It works in step by step manner as shown in the block diagram in fig 2. The first step in the cycle is to sense the color intensities of the image, named as thresholding [F. SHAFAIT, D. K. San Jose, 2008], and converts the image into binary images. Second step is to do the connected component analysis [SMITH, R, 2007] of the image, which does the task of extracting character outlines. This step is the main process of this cycle as it does the text recognition of image with white text and black background of the image. Tesseract uses these cycles to process the input image. After this the outlines extracted from image are converted into Blobs (Binary Large Objects). It is then organized as lines and regions and further analysis is done for some fixed area [SMITH, R, 2007]. After extraction, the extracted components are chopped into words and delimited with spaces. Recognition of text then starts which is a two pass process. As shown in fig 2, the first part is when an attempt to recognize each word is made. Each satisfactory word is accepted and second pass is commenced to gather remaining words. This brings in the role of adaptive classifier which will classify text in more accurate manner. The adaptive classifier needs to be trained beforehand to work accurately. When the classifier receives some data, it has to resolve the issues and locates the text.

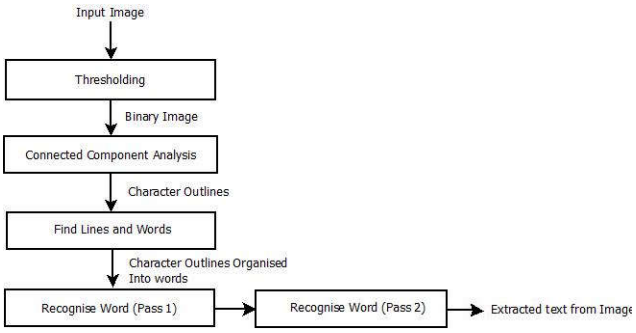


Figure 2 Tesseract Flow

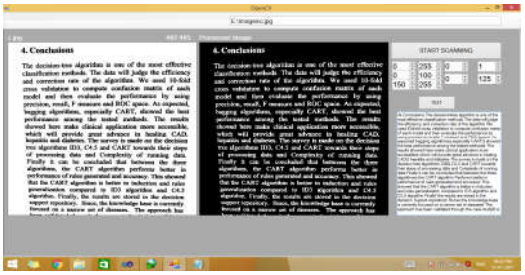
Character Database

Image that is fed in the database are segmented and stored as binary data. This binary data is converted into text stream and the list of all characters that are fed to the machine for learning are present in this database. The database contains different fonts and font-sizes. The query word provided by the user and the list of words extracted from the image databases are matched, and the most relevant and morphologically correct results are obtained.

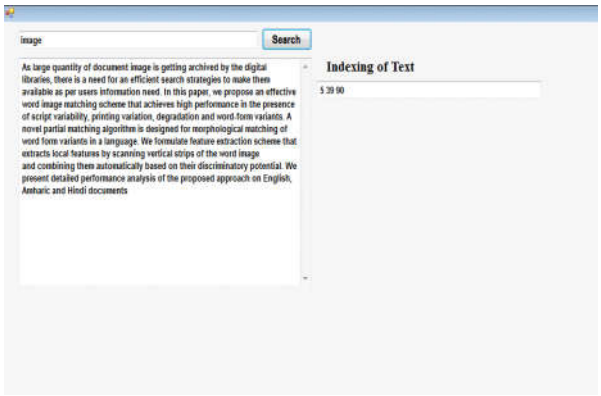
Retrieval

During the retrieval phase, a query word image is presented to the system. The word is segmented and features are extracted of each character. The extracted character is then compared with the characters in the text stream. Once the closest feature is determined, the index file associated with the location of the character in the database is parsed to retrieve all the documents containing the occurrences of the query word. The process is repeated for all the characters in the query word and finally the retrieval results are merged to keep only those documents which contain the complete query word. The retrieval results along with the query words highlighted are presented to the user.

- Output
- *Text Extraction*



- *Text Retrieval*



Applications

Word Searching

One of the applications of the proposed system is word image matching [Seema Yadav, Dr. Sudhir Sawarkar, 2009]. Searching/locating a user-specified keyword in image format documents has been a topic of interest for many years. It has its practical value for document information retrieval. For example, by using this technique, the user can locate a specified word in document images without any prior need for the images to be OCR-processed.

Banking

The uses of image text recognition vary across different fields [Mr. Pratik Madhukar Manwatkar, Mr. Shashank H. Yadav, 2015]. One widely known application is in banking, it is used to process checks without human involvement. A check can be inserted into a machine, the writing on it is scanned instantly, and the correct amount of money is transferred. This technology has nearly been perfected for printed checks, and is fairly accurate for handwritten checks as well, though it occasionally requires manual confirmation.

Legal

In the legal industry, there has also been a significant movement to digitize paper documents [Mr. Pratik Madhukar Manwatkar, Mr. Shashank H. Yadav, 2015]. In order to save space and eliminate the need to sift through boxes of paper files, documents are being scanned and entered into computer databases. Image text recognition further simplifies the process by making documents text-searchable, so that they are easier to locate. Legal professionals now have fast, easy access to use library of documents in electronic format, which they can find simply by typing in a few keywords.

Healthcare

Healthcare also makes use of image text recognition technology to process paperwork [Mr. Pratik Madhukar Manwatkar, Mr. Shashank H. Yadav, 2015]. Healthcare professional always have to deal with large volumes of forms for each patient, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. By using image recognition technology they are able to extract information from forms and put into database, so that every patient's data is promptly recorded. As a result, healthcare providers can focus on delivering best possible service to every patient.

CONCLUSION

The proposed system works on text retrieval from scanned document images. Document image retrieval without OCR has its practical value, but it is also a challenging problem. Current information retrieval systems do not properly retrieve the query due to poor quality of image consisting of noise. To enhance the quality of image we pre-process the image and this image is fed to tesseract for segmentation. Tesseract is currently the best open source tool for extraction of text from image documents due to which we will be getting best retrieval of user query of image documents. Constant updated trained data helps in increased accuracy of the system. We propose an efficient and scalable system which is capable of handling large volumes of data and retrieve the word efficiently and accurately.

References

1. Y. He, Z. Jiang, B. Liu, and H. Zhao, "Content-Based Indexing and Retrieval Method of Chinese Document Images," Prof Fifth Int'l Conf Document Analysis and Recognition (ICDAR '99), pp. 1999.
2. Seema Yadav, Dr. Sudhir Sawarkar, Retrieval Of Information In Document Image Databases Using Partial Word Image Matching Technique, 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.
3. Raashid Hussain, Haris Ahmad Khan, Imran Siddiqi, Khurram Khurshid, Asif Masood, Keyword based Information Retrieval System for Urdu Document Images 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems.
4. Pratiksha Jain, Neha Chopra, Vaishali Gupta, Automatic License Plate Recognition using OpenCV, *International Journal of Computer Applications Technology and Research* Volume 3– Issue 12, 756 - 761, 2014.
5. Million Meshesha and C. V. Jawahar, Matching word images for content-based retrieval from printed document images, Proceeding *of the International Journal on Pattern Recognition*, *DOI 10.1007/s10032-008-0067-3*, *2008.*
6. Pramod Sankar K, R Manmatha and C V Jawahar - Large Scale Document Image Retrieval by Automatic Word Annotation *International Journal on Document Analysis and Recognition (IJDAR): Volume 17, Issue 1(2014), Page 1-17.*
7. SMITH, R. 2007. An Overview of the Tesseract OCR Engine. In proceedings of Document analysis and Recognition, ICDAR 2007. IEEE Ninth International Conference.
8. D. Doermann, "The Indexing and Retrieval of Document Images:A Survey," Computer Vision and Image Understanding, vol. 70, no. 3,pp. 287-298, 1998.
9. Mr. Pratik Madhukar Manwatkar, Mr. Shashank H. Yadav, Text Recognition from Images, *IEEE Sponsored 2nd International Conference on Innovations in Information, Embedded and Communication systems (ICIIECS)2015.*

10. D. Doermann, "The Indexing and Retrieval of Document Images: A Survey," Computer Vision and Image Understanding, vol. 70, no. 3,pp. 287-298, 1998.
11. Mr. Pratik Madhukar Manwatkar, Mr. Shashank H. Yadav, Text Recognition from Images, *IEEE Sponsored 2nd International Conference on Innovations in Information, Embedded and Communication systems (ICIIECS)2015*.
12. Tin Kam Ho, Jonathan J. Hull and Sargur N. Srihari, A Word Shape Analysis Approach to Lexicon Based Word Recognition, Article in Pattern Recognition Letters, November 1992.
13. X. Tong and D. A. Evans, (1996) "A statistical approach to automatic OCR error correction in context," in Fourth Workshop on Very Large Corpora (WVLC-96), pp. 88–100.
14. Ankit Sharma, Dipti R Chaudhary, Character Recognition Using Neural Network, International *Journal of Engineering Trends and Technology* (IJETT) - Volume4 Issue4- April 2013.
15. F. SHAFAIT, D. K. San Jose, CA: s.n., 2008.Efficient Implementation of Local Adaptive Thresholding Techniques Using Integral Images In Document Recognition and Retrieval XV, S&T/SPIE Annual Symposium on Electronic Imaging.

****

How to cite this article:

Seema Yadav *et al*.2017, Information Retrieval From Image Databases. *Int J Recent Sci Res.* 8(1), pp. 15279-15283.