Image Content from Scanned Documents

Yesh Rana[1], Prathamesh Chougule[2], Mayur Budhbhatti[3], Sanket Shidhore[4], Prof. Abirami Sivaprasad[5]
Department of Information Technology
[1, 2, 3, 4, 5] Shah & Anchor Kutchhi Engineering College ,Mumbai University, India

Abstract*-As large quantity of document images is getting archived by the digital libraries, there is a need for an efficient search strategies to make them available as per users information need. For their retrieval, it is important 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. Thus there is a need of character recognition mechanisms to perform Document Image Analysis (DIA) which transforms documents in paper format to electronic format. In this paper we have discuss method for text recognition from images. The objective of this paper is to recognition of text from image and retrieval of user query based on user interest.*

Keywords- Document image analysis, segmentation,Tesseract Engine, OpenCV, Indexing, Preprocessing.

I. INTRODUCTION

With storage becoming cheaper and imaging devices becoming increasingly 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. 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 errors may include edition operations such as characters substitution, deletion, and insertion.

The goal of this research is to design an IR method to search large textual databases and return the documents that the system considers relevant to the user's query. In particular, we will take into account the possible recognition errors using the retrieval process[7].

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. Preprocessing 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.

1.1 RELATED WORK

Rapid access to information is a major advancement obtained through digital technology where information is digitized and made available online to all stakeholders. However, there is still a huge document base in printed form in libraries and in order to make these accessible to all, digital libraries play a vital role. The concept of a digital library is not limited to mere scanning of books and documents. These scanned documents need to be complemented by an information retrieval system allowing 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. 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, [10] Ho et al proposed a word recognition method based on word shape analysis without character segmentation and recognition. 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. In 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. These approaches can reduce the overall OCR error rates for the frequent words of the language, but it is likely to corrupt correctly recognized words which are not in the dictionary, for instance names and places. As an alternative Tong et al proposed to use the context of the text itself to correct misrecognized words. 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. In, the authors demonstrate that best results were achieved using edge-based text detection compared to mathematical morphology and colour-based character extraction.

II. 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. Preprocessing 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.

OpenCV, is basically a library of functions written in C/C++. Here we use OpenCV library files integrated in Visual Studios. Programs written in OpenCV 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 OpenCV is fast when it comes to speed of execution. OpenCV is based on C. There is better memory management here because every time a chunk of memory is allocated, it has to be released again. If you have a loop in your code where you allocate a chunk of memory in that loop and forget release it afterwards, you will get what is called a “leak”. This is where the program will use a growing amount of memory until it crashes from no remaining memory. Hence

	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

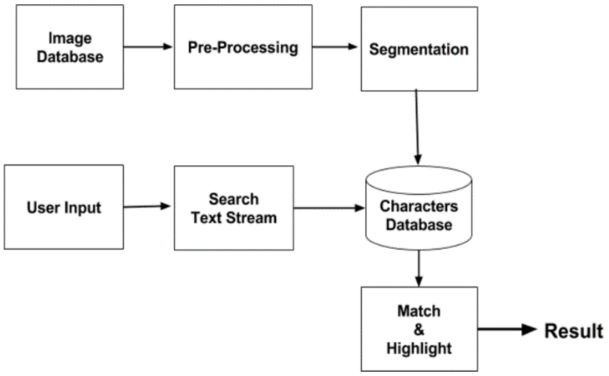

Figure 3: System Diagram for searching user specified word in document image

2.1 USER INTERACTON

Here the user will enter query of interest. After processing the query, the documents consisting user query word are listed and user can access the documents of interest.

2.2 PRE-PROCESSING

Pre-processing involves the set of all documents are gathered and passed to the word extraction phases in which all words are extracted. Preprocessing techniques are required in 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 processing color images is computationally high. Such images may also contain non-uniform background and/or watermarks making it difficult to extract the document text from the image without performing some kind of preprocessing, therefore; the desired result from preprocessing is a binary image containing text only. Thus, to achieve this, several steps are needed, first, some image enhancement techniques to remove noise or correct the contrast in the image, second, thresholding to remove the background containing any scenes, watermarks and/or noise, third, character segmentation to separate characters from each other and, finally, morphological processing to enhance the characters in cases where thresholding and/or other preprocessing techniques eroded parts of the characters or added pixels to them.

2.3 SEGMENTATION

The segmentation is the most important process in text recognition. Segmentation is done to make the separation between the individual characters of an image and is one of the most important phases in this project. The performance of this project is depending on segmentation. Images will be fed to Tesseract Engine where it performs character segmentation and that 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 constituent of the script, which are certainly characters. This is needed because our classifier recognizes these characters only.

2.4 WORD SPOTIING – TESSERACT ENGINE

Tesseract is an example based system. This makes it efficient and flexible. By example based systems we mean that the engine works on a set of example rules defined in the system and results depend on this data. So in simpler words to get good results we need to define these set of rules properly 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 OCR is an elegant engine with various layers. It works in step by step manner as shown in the block diagram in fig 4. The first step in the cycle is to sense the color intensities of the image, named as adaptive thresholding [9], and converts the image into binary images. Second step is to do the connected component analysis [7] of the image, which does the task of extracting character outlines. This step is the main process of this cycle as it does the OCR of image with white text and black rest of the image. Tesseract was probably the first [7] to use these cycles to process the input image. After this the outlines extracted from image are converted into Blobs (Binary Long Objects). It is then organized as lines and regions and further analysis is for some fixed area [7]. After extraction the extracted components are chopped into words and delimited with spaces. Recognition in text then starts which is a two pass process. As shown in fig 1, the first part is when attempt to recognize each word is made. Each satisfactory word is accepted and second pass is started to gather remaining words. This brings in the role of adaptive classifier. The adaptive classifier then 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 assign the proper place of the text. More details regarding every step is available at [7].

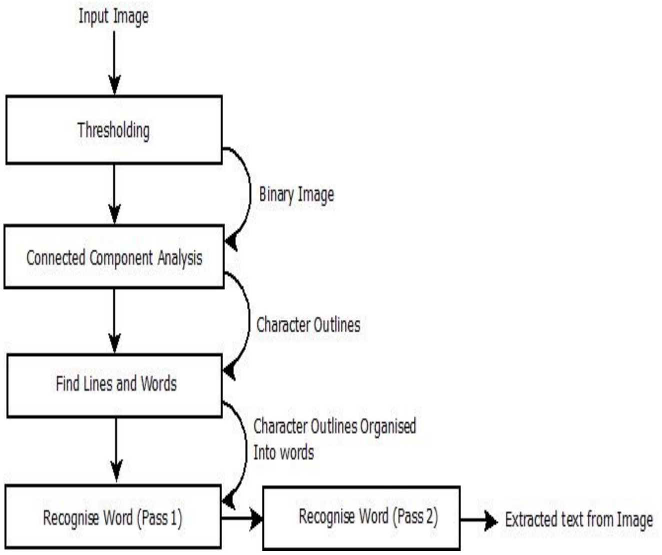

Figure 4: Tesseract Flow

2.5 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.

2.6 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. A query 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 on 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.

III. FEASIBILITY STUDY

Document Image Processing (DIP) is utilized to automatically convert the digital images of documents to the machine-readable text format using Optical Character Recognition (OCR) technology. Although Optical Character Recognition (OCR) scanning technology has increased rapidly over the years, there are, however, limitations in regards to the source materials and character formatting. Most document formatting is lost during text scanning, except for paragraph marks and tab stops. Sometimes bold, italics and underline remain unrecognized by OCR systems. Interactive manual correction/proofreading of OCR results is usually unavoidable in most DIP systems. Source materials that often cause issues are small text, blurry copies and unusual or script-type fonts. OCR is not a cost effective and a practical way to process a large number of paper documents.

3.1 Operational Feasibility

Two types that we consider are firstly Efficient Pre-Processing of Dynamically available Documents and secondly Segmentation of heavily touching characters.

3.2 Market Feasibility

Market Feasibility involves Advancement in currently available tools and cost effectiveness of the system.

IV. APPLICATIONS

4.1 Word Searching

We now focus on the application of our proposed word image matching method. 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[2].

4.2 Banking

The uses of image text recognition vary across different fields. 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. Overall, this reduces wait times in many banks[9].

4.3 Legal

In the legal industry, there has also been a significant movement to digitize paper documents. 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 and work with once in the database. 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[9].

4.4 Healthcare

Healthcare also use of image text recognition technology to process paperwork. 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[9].

V. CONCLUSION

Document images have become a popular information source in our modern society, and information retrieval in document image databases is an important topic in knowledge and data engineering research. 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. We propose an efficient and scalable system which is capable of handling large volumes 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Y.FATAICHA, M.CHERIET, J. Y. NIE, and C. Y. SUEN, Information retrieval based on OCR errors in scanned documents.

[7] D. Doermann, “The Indexing and Retrieval of Document Images:A Survey,” Computer Vision and Image Understanding, vol. 70, no. 3,pp. 287-298, 1998.

[8] 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.

[9] 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.