Word Matching and Retrieval from Images

Seema Yadav
Assistant Professor, Department of Information Technology,
K.J. Somaiya Institute of Engineering and Information Technology,
University of Mumbai, Maharashtra, India.
syadav@somaiya.edu

Priya Bhanushali
B.E student, Department of Information Technology,
K.J. Somaiya Institute of Engineering and Information Technology,
University of Mumbai, Maharashtra, India.
pab@somaiya.edu

Saurabhkumar Jain
B.E student, Department of Information Technology,
K.J. Somaiya Institute of Engineering and Information Technology,
University of Mumbai, Maharashtra, India.
saurabhkumar.j@somaiya.edu

Tejinder Kaur
B.E student, Department of Information Technology,
K.J. Somaiya Institute of Engineering and Information Technology,
University of Mumbai, Maharashtra, India.
t.rehan@somaiya.edu

Abstract—As vast amount of digital image data is stored by the advanced libraries, there is a requirement for an efficient query word searching methodologies which can make them accessible according to user's requirement. For their accurate retrieval, it is essential to understand their contents. Present technologies for optical character recognition (OCR) and image document analysis do not handle such documents adequately because of the recognition errors. Due to the problems faced by traditional OCR during recognition, computer is unable to extract the textual characters properly after scanning them. In this paper, we propose an effective word extraction and matching scheme from image documents that achieves high performance, even in the presence of noise in the image, degradation and font-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 system is able to properly handle images with different font styles, size and heavily touching characters. From the experimental results on the various image formats 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.

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

I. INTRODUCTION

With storage getting to be distinctly available and imaging devices becoming 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 converts scanned document images into text[5]. Due to the distortion and the poor contrast in the images, many extraction attributes must be acquired to differentiate text from complex textual image. Secondly, it is tough 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[14].

The objective of this system is to plan an IR strategy to extract extensive character databases and give back the archives that framework considers applicable to the user's query. Specifically, we will also consider the possible text recognition errors using the retrieval process[14].

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.

II. 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 keeping in mind the main objective is 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[3]. 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.

Gur et al. [21] has discussed some problems in text recognition and retrieval. Automated optical character recognition (OCR) tools do not supply a complete solution and in most cases human inspection is required. They suggest a novel text recognition algorithm based on usage of fuzzy logic rules relying on statistical data of the analyzed font. The new approach combines letter statistics and correlation coefficients in a set of fuzzy based rules, enabling the recognition of distorted letters that may not be retrieved otherwise.

Badawy, W. et al. [22] has discussed the Automatic license plate recognition (ALPR) is the extraction of vehicle license plate information from an image or a sequence of images. The extracted information can be used with or without a database in many applications, such as electronic payment systems (toll payment, parking fee payment), and freeway and arterial monitoring systems for traffic surveillance. The ALPR uses either a color, black and white, or infrared camera to take images.

Jawahar et al. [23] has proposed a recognition scheme for the Indian script of Devanagari. They used approach does not require word to character segmentation, which is one of the most common reason for high word error rate. They have been reported a reduction of more than 20% in word error rate and over 9% reduction in character error rate while comparing with the best available OCR system.

Ntirogiannis et al. [24] has studied that the document image binarization is of great importance in the document image analysis and recognition pipeline since it affects further stages of the recognition process. The evaluation of a binarization method aids in studying its algorithmic behavior, as well as verifying its effectiveness, by providing qualitative and quantitative indication of its performance. They proposed a pixel-based binarization evaluation methodology for historical handwritten/machine-printed document images.

Malakar et al. [8] has described that extraction of text lines from document images is one of the important steps in the process of an Optical Character Recognition (OCR) system. In case of handwritten document images, presence of skewed, touching or overlapping text line makes this process a real challenge to the researcher.

III. 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 textual characters from images. Tesseract efficiently handles extraction of text from white on black images[7]. 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 a trainable classifier. 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 manage the issue of heavily touching characters.

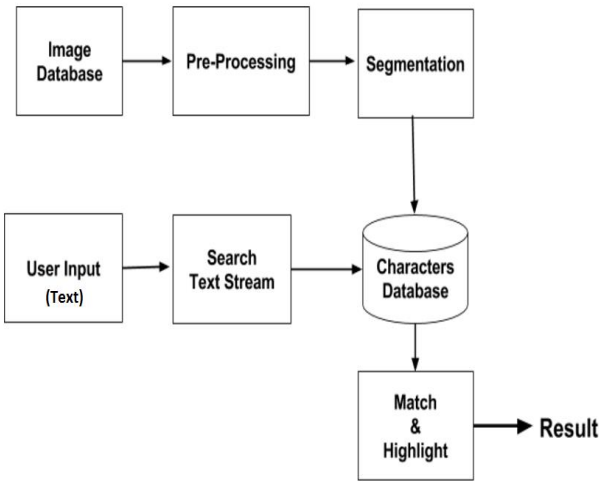


Figure 1: System Diagram for retrieving user specified keyword in image database

As image processing is one of the base domain of the system. So OpenCV library is integrated in the system for processing of image documents. OpenCV, is basically a library of functions written in C/C++[17]. 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.

Table 1 : Index Scores comparison between OpenCV 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**
Programming Skills	**3**	**8**
Help and sample code	**8**	**9**
Debugging	**9**	**5**
Total	**59**	**71**

IV. LIST OF MODULES

A. 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.

B. 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 textual characters. In character recognition method most of the applications use grey level or binary images since dealing with colored images requires more processing. Such images may also contain noise and watermarking which makes it difficult to extract the textual characters from the image without performing some kind of pre-processing. Therefore, the desired result from pre-processing is a binary image containing noise free text. To achieve this, several steps are performed. Firstly, some image enhancement techniques are used to remove noise or to correct the contrast in the image[12]. Secondly, adaptive thresholding is carried out to remove the background containing any scenes,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 textual characters. In character recognition method most of the applications use grey level or binary images since dealing with colored images requires more processing. Such images may also contain noise and watermarking which makes it difficult to extract the textual characters from the image without performing some kind of pre-processing. Therefore, the desired result from pre-processing is a binary image containing noise free text. To achieve this, several steps are performed. Firstly, some image enhancement techniques are used to remove noise or to correct the contrast in the image[12]. Secondly, adaptive thresholding is carried out to remove the background containing any scenes, 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.

D. 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 [13], 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 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 [7]. 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.

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 textual characters. In character recognition method most of the applications use grey level or binary images since dealing with colored images requires more processing. Such images may also contain noise and watermarking which makes it difficult to extract the textual characters from the image without performing some kind of pre-processing. Therefore, the desired result from pre-processing is a binary image containing noise free text. To achieve this, several steps are performed. Firstly, some image enhancement techniques are used to remove noise or to correct the contrast in the image[12]. Secondly, adaptive thresholding is carried out to remove the background containing any scenes, 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.

D. 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 [13], 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 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 [7]. 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.

E. 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.

F. 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.

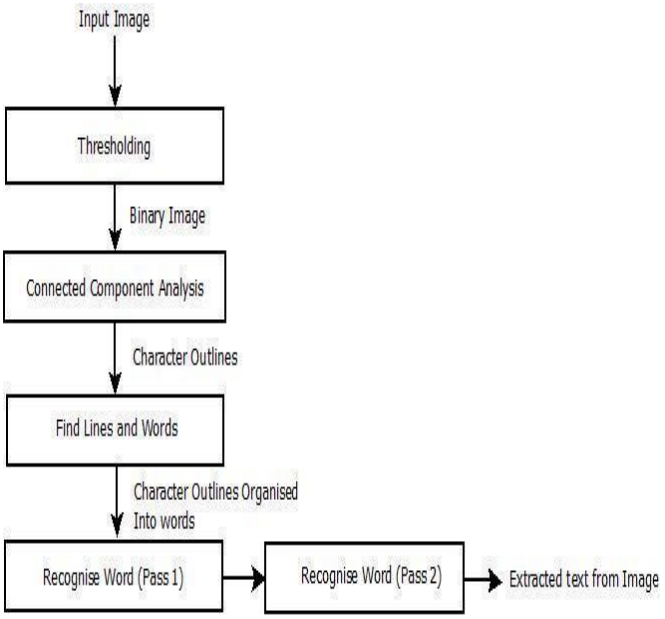
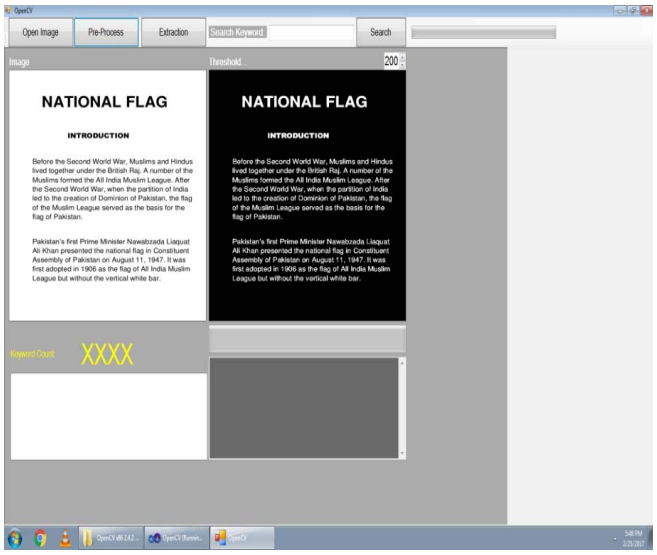


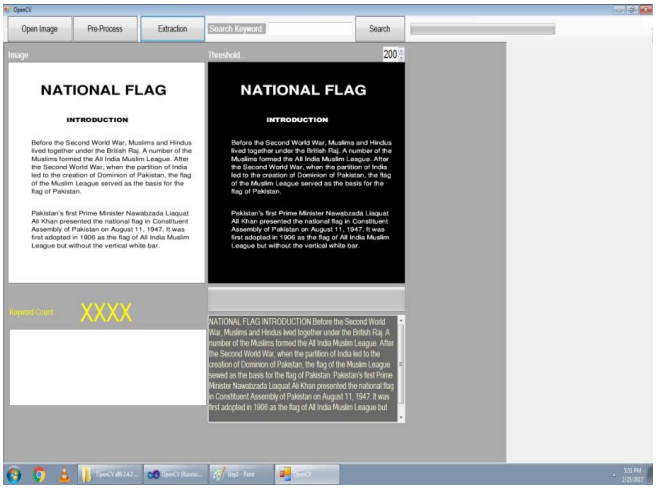
Figure 2: Tesseract Segmentation Flow

V. OUTPUT

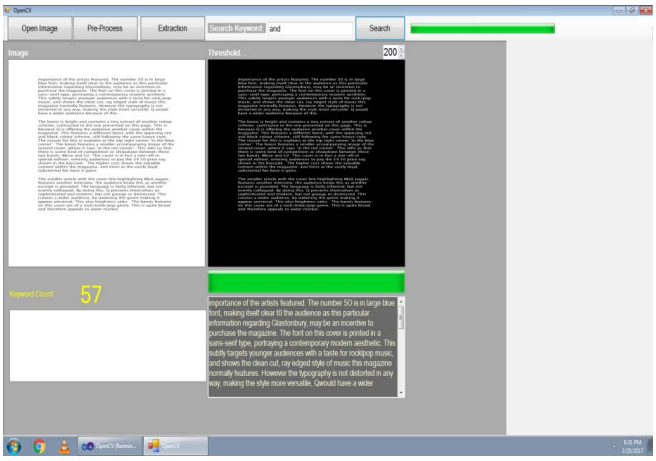
STEP I: Pre-processing



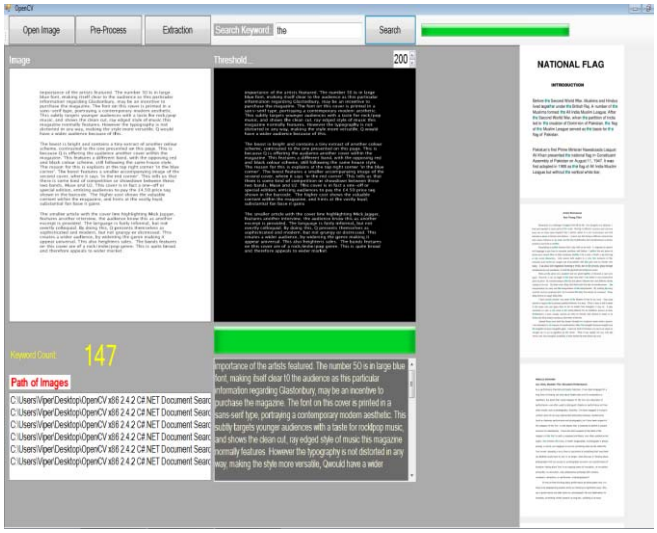
STEP II: Text Extraction



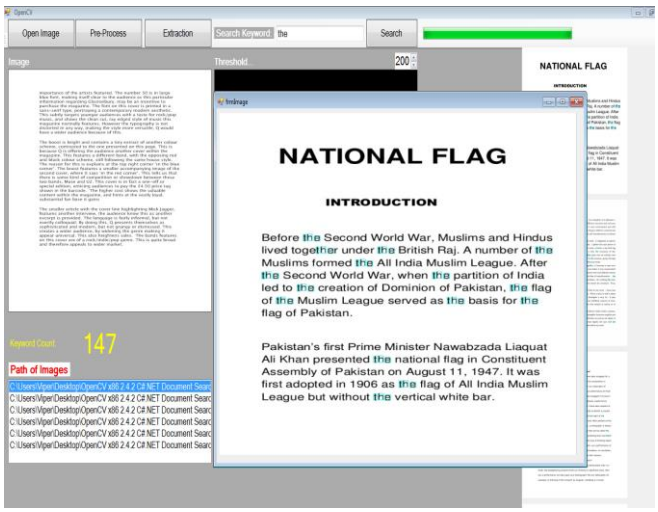
STEP III: Keyword Matching and Count



STEP IV: Listing of File Path



STEP V: Keyword Retrieval and Highlighting



VI. 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.

VII. 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] Text Detection and Recognition in Imagery: A Survey Qixiang Ye; David Doermann IEEE Transactions on Pattern Analysis and Machine Intelligence.Year: 2015, Volume: 37, Issue: 7

[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] 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.

[11] 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.

[12] Ankit Sharma, Dipti R Chaudhary, Character Recognition Using Neural Network, International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue4- April 2013.

[13] 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.

[14] Y. Fataicha. "Information Retrieval Based on OCR Errors in Scanned Documents", 2003 Conference on Computer Vision and Pattern Recognition Workshop, 06/2003

[15] Wemhoener, David, Ismet Zeki Yalniz, and R. Manmatha. "Creating an Improved Version Using Noisy OCR from Multiple Editions", 2013 12th International Conference on Document Analysis and Recognition, 2013.

[16] Ikica, A., and P. Peer. "An improved edge profile based method for text detection in images of natural scenes", 2011 IEEE EUROCON - International Conference on Computer as a Tool, 2011.

[17] Manwatkar, Pratik Madhukar, and Shashank H. Yadav. "Text recognition from images", 2015 International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS), 2015.

[18] Vinumol, K.P., Ashsish Chowdhury, Radhika Kambam, and V. Muralidharan. "Augmented Reality Based Interactive TextBook: An Assistive Technology for Students with Learning Disability", 2013 XV Symposium on Virtual and Augmented Reality, 2013.

[19] Tan, Chew Lim, Xi Zhang, and Linlin Li. "Image Based Retrieval and Keyword Spotting in Documents", Handbook of Document Image Processing and Recognition, 2014.

[20] Yje Lu. "Information retrieval in document image databases", IEEE Transactions on Knowledge and Data Engineering, 11/2004

[21] Retrieval of Rashi Semi-cursive Handwriting via Fuzzy Logic Eran Gur; Zeev Zelavsk 2012 International Conference on Frontiers in Handwriting Recognition Year: 2012

[22] Automatic License Plate Recognition (ALPR): A State-of-the-Art Review-Shan Du; Mahmoud Ibrahim; Mohamed Shehata; Wael Badawy IEEE Transactions on Circuits and Systems for Video Technology Year: 2013, Volume: 23, Issue: 2

[23] Recognition of printed Devanagari text using BLSTM Neural Network Naveen Sankaran; C. V Jawahar Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)Year: 2012

[24] Performance Evaluation Methodology for Historical Document Image Binarization Konstantinos Ntirogiannis; Basilis Gatos; Ioannis Pratikakis IEEE Transactions on Image Processing Year: 2013, Volume: 22, Issue: 2

[25] Information Retrieval From Image Databases, Seema Yadav, Saurabhkumar Jain, Priya Bhanushali and Tejinder Kaur, International Journal of Recent Scientific Research Vol. 8, Issue, 1, pp. 15729-15283, January, 2017