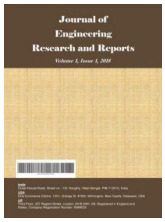

Journal of Engineering Research and Reports

2(1): 1-10, 2018; Article no.JERR.43002

Reading and Interpreting Machine Printed Text in Camera-Captured Document Images

V. J. Rehna[1*], Abid Siddique[1] and Sreenivas Naik[1]

[1]Department of Engineering, Ibri College of Technology, Ibri, Oman.

Authors' contributions

This work was carried out in collaboration between all authors. The first and corresponding author RVJ designed the study, wrote the protocol and penned the first draft of the manuscript. The second author AS conducted the literature searches and performed the statistical analysis. The third author, SN managed the analyses of the study. Both the second and third authors edited the manuscript. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JERR/2018/v2i19904
Editor(s):
(1) P. Elangovan, Associate Professor, Department of EEE, Sreenivasa Institute of Technology and Management Studies, Chittoor, Andhra Pradesh, India.
Reviewers:
(1) Zlatin Zlatev, Trakia University, Bulgaria.
(2) K. Karthick, GMR Institute of Technology, Razam, India.
(3) S. Rohith, Nagarjuna College of Engineering and Technology, Viswesvarya Technological University, India.
Complete Peer review History: http://www.sciencedomain.org/review-history/26071

Original Research Article

Received 19th May 2018
Accepted 26th July 2018
Published 3rd September 2018

ABSTRACT

Aims: To introduce a cost-effective tool for reading and interpreting machine printed text in document images and save as computer-processable codes.
Study Design: In this work, emphasize is given on extracting uppercase & lowercase letters and numerals from document images by the technique of segmentation and feature extraction using MATLAB Image Processing toolbox.
Place and Duration of Study: Department of Engineering, Ibri College of Technology, between September 2017 and May 2018.
Methodology: Necessary information about existing algorithms on character recognition is collected by review of relevant literature available in journals, books, manuals and related documents. Suitable architecture and novel algorithm for a simple, low cost, low complexity, highly accurate system is developed as per the specifications and reviewed literature. Functionality of the design is verified using simulation software MATLAB.
Results: The proposed method can extract characters from document image (which may be

**Corresponding author: E-mail: rehnavj09@gmail.com, rehnavj@ibrict.edu.om;*

scanned or camera captured) of any font size, colour, space and can be rewritten in an editable window like Notepad, WordPad where the characters can even be edited; thus, improving accuracy and hence, saves time.
Conclusion: This algorithm gives promising results that have been obtained on a number of images in which almost all characters are retrieved. It also gives 90 percent accuracy for all printed characters.

Keywords: Document image analysis; optical character recognition; segmentation; feature extraction; classification.

1. INTRODUCTION

Making a machine or computer able to perform tasks such as reading a document and writing it down, just as human would, is being materialized. A document image contains different kinds of information such as texts, pictures and graphics i.e., line-drawings and sketches and are made by scanning journals, historical documents, degraded documents, handwritten texts, printed documents, multi-color book covers, newspapers etc. There are many challenges which are faced while processing scanned documents including low contrast, low resolution, color bleeding, complex background and unknown text color, size, position, orientation, layout etc. The objective of document image analysis is to recognize the text and graphic components in images and extract the intended information in a similar way as human would. Some methods include recognizing the text by OCR (Optical Character Recognition) by determining the skew, columns, text lines and words. Text in images contains meaningful and useful information which can be used to fully understand the contents of the images. Text extracted from images play an important role in document analysis, vehicle plate detection, video content analysis, document retrieval and also helps the blind and visually impaired users. Character extraction is the extraction of texts from document images and analysis of the same and is one of the key tasks in document image analysis and subtask of page segmentation. The process mainly involves Segmentation, Feature Extraction and Classification. Extraction of characters from documents in which character is embedded in complex coloured document image is a very challenging problem.

There are many potential uses of character extraction including Banking, Business, and Healthcare. Character extraction from images finds many useful applications in document processing, analysis of technical papers with tables, maps, charts, and electric circuits, identification of parts in industrial automation, content-based image/video retrieval from image / video databases, educational & training videos and TV programs such as news containing mixed text-picture-graphics regions. The other applications include document retrieval, camera based document analysis etc.

Character / text extraction is done in order to speed up the data entry process. One of the main reasons for employing this method is the reduction in the manual data entry errors. The current automatic data retrieval system prevailing in the market uses indirect method of reading data from the document. For e.g. Watermark based system to read confidential data and bar code system to read the price of goods in the retail commercial outlets. In this work, a system of direct reading of text from documents is introduced which can be used as an alternative to the existing technology. The main objective is to perform character extraction, recognition and interpretation of texts from document images. The result is then exported to spread sheets where it can be edited.

The paper is organized as follows: section 2 deals with the basics of document image analysis, section 3 reviews relevant literature, section 4 describes the methodology of implementation, section 5 deals with the flow diagram, followed by results and discussion in section 6. Future prospects are discussed in section 7 and finally conclusion of the paper is presented in section 8.

2. DOCUMENT IMAGE ANALYSIS

Document Image analysis (DIA) is the theory and practice of recovering the symbol structures of digital images scanned from paper or produced by computer. Studying the content and structure of the documents, identifying and naming the components of some class of documents, specifying their interrelationships and naming their properties is known as document image

analysis. Data capture of documents by optical scanning or by digital video yields a file of picture elements or pixels that is the raw input to document analysis. The first step in document analysis is to perform processing on this image to prepare it for further analysis. Such processing includes thresholding to reduce a gray-scale or color image to a binary image [1], reduction of noise to reduce extraneous data, and thinning and region detection to enable easier subsequent detection of pertinent features and objects of interest. Document image analysis and recognition (DIAR) is a research field that has its roots in the first Optical Character Recognition (OCR) systems.

The objective of document image analysis is to recognize the text and graphics components in images of documents, and to extract the intended information as a human would. Two categories of document image analysis can be defined. Textual processing deals with the text components of a document image. Some tasks here are determining the skew (any tilt at which the document may have been scanned into the computer), finding columns, paragraphs, text lines, and words, and finally recognizing the text (and possibly its attributes such as size, font etc.) by optical character recognition (OCR). Graphics processing deals with the non-textual line and symbol components that make up line diagrams, delimiting straight lines between text sections, company logos etc. Pictures are a third major component of documents, but except for recognizing their location on a page, further analysis of these is usually the task of other image processing and machine vision techniques [2].

There are two main types of analysis that are applied to text in documents. One is optical character recognition (OCR) to derive the meaning of the characters and words from their bit-mapped images, and the other is page-layout analysis to determine the formatting of the text, and from that to derive meaning associated with the positional and functional blocks (titles, subtitles, bodies of text, footnotes etc.) in which the text is located. Depending on the arrangement of these text blocks, a page of text may be a title page of a paper, a table of contents of a journal, a business form, or part of an email. OCR and page layout analysis may be performed separately, or the results from one analysis may be used to aid or correct the other.

OCR deals with automatic conversion of scanned images of machine printed or handwritten documents into computer processable codes. The goal of Optical Character Recognition (OCR) is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters. The process of character extraction [3] involves several steps namely segmentation, feature extraction and classification. Segmentation is the process of partitioning a digital image into multiple segments. Segmentation is divided into two levels. In the first level, text and graphics are separated and are sent for subsequent processing. In the second level; the rows, columns, paragraphs and words are recognized. Segmentation is of two types, namely implicit segmentation and explicit segmentation. In implicit segmentation words are recognized and in explicit segmentation individual characters are recognized [4]. In this work, explicit segmentation method is employed.

The attributes of the image are recovered in feature extraction stage [5]. The number of lines, line spacing and number of crossings are noted here. Classification is the process of defining undefined objects. There are two approaches to Classification namely, Statistical Classification and Structural Classification. In Statistical Classification approach, character image patterns are represented by points. In Structural Classification approach; strokes, cross points and end points are noted [6]. Structural Classification approach is adopted in this work [7].

3. LITERATURE REVIEW

A large number of techniques have been used for text extraction of document images. Also, many research efforts have been made to detect text regions in natural scene images. The text detection algorithms can be roughly divided into two categories: rule-based algorithms and feature-based algorithms. In 1996, Suen et al. [8] proposed a method for extracting text strings from the colour printed document in a 24-bit true colour image. The processing is time consuming due to the very large amount of data in 24-bit colour image. In 1999, Sobottka et al. [9] proposed a method based on top-down and bottom-up analysis for text location and identification on coloured book and journals. In 2001, Yuan et al. [10] proposed a method using edge information to extract text from the grey-scale document images, i.e. heavy noise infected newspaper. In 2004, Raju et al. [11] formulated a texture based and connected component

analysis approach for text extraction from the document image. In 2005, Shi et al. [12] proposed an algorithm using adaptive local connectivity map (ALCM) for text extraction from the complex handwritten historical document. In 2006, Qiao et al. [13] suggested a Gabor filter based method to extract text from document images. This method was also used to extract text from ancient damaged documents of the 18th and 19th century. In 2009, Grover et al. [14] proposed edge based feature for detection of text embedded in complex coloured document images. Audithan et al. [15] described an effective method for text extraction from document images using Haar discrete wavelet transform (DWT). 2D-Haar DWT detects all the horizontal, vertical and diagonal edges. The non-text edges are further removed by using thresholding technique. Then, morphological dilation operators are used to connecting the text edges in each detail component. Further, logical AND operator is used to removing the remaining non-text regions. In 2010, Hoang et al. [16] proposed a text extraction method from graphical document images by applying morphological component analysis (MCA). MCA allows the separation of features contained in an image by promoting sparse representation of these features in two chosen dictionaries. Curvelet transform and Undecimated Wavelet Transform (UWT) dictionary is used for graphics and text respectively. Li et al. [17] proposed an effective interpolation-based resolution enhancement (RE) algorithm for low-resolution document images. Malakar et al. [18] proposed a spiral run length smearing algorithm (SRLSA) for text extraction from handwritten document pages.

4. METHODOLOGY

Text line segmentation is a major component for document image analysis. Text in documents depend upon various factors such as language, styles, font, sizes, color, background, orientation, fluctuating text lines, crossing or touching text lines. Character extraction is mainly done to ease the process of machine reading. Detection of character from documents in which text is embedded in complex coloured document images is a very challenging task. It involves segmentation, feature extraction, classification, recognition and interpretation of the character [19]. Execution is divided into two main stages namely training and testing as shown in Fig. 1. Training data is the data which is used to create the database and test data is the input image under consideration. These images are made to undergo following steps.

4.1 Pre-processing

The image resulting from the scanning process may contain a certain amount of noise and hence, pre-processing is required. In the pre-processing stage, input colour image (a camera captured or scanned document image) is converted to a grey-scale image. This image is then filtered by using a median filter to remove noise. Next, Sobel edge detector is applied to extract strong edges and separate the background from the object. Edges are mapped to connect edges of the same object. Order static filter is used to merge the neighbouring edges to form a single text region. Some of the non-text regions are removed based on the structural property of the text, i.e. specific height and width of the text. By applying threshold and median filter, the image is binarized to clarify the text and noise is removed respectively.

4.2 Feature extraction

The relevant information about the characters is extracted from the pre-processed image in the feature extraction stage by partitioning the image into multiple segments in the segmentation stage. The result of segmentation is image attributes which are made to undergo a subsequent process of matching. Errors in the segmentation process may result in confusion between text and graphics or between texts and noise.

4.3 Classification

These features/attributes are then interpreted and classified accordingly in the classification stage. Structural classification approach is adopted here. Incorrect classification may result due to poor design of the classifier.

4.4 Database Creation

The images of uppercase letters A to Z and lowercase letters a to z, numerals 0 to 9 are recorded and are called sample images. They are resized to the dimensions of the extracted characters, recognized and assigned a label depending upon its features.

The extracted features are then stored in a database for further processing. In the identification stage, features of the extracted

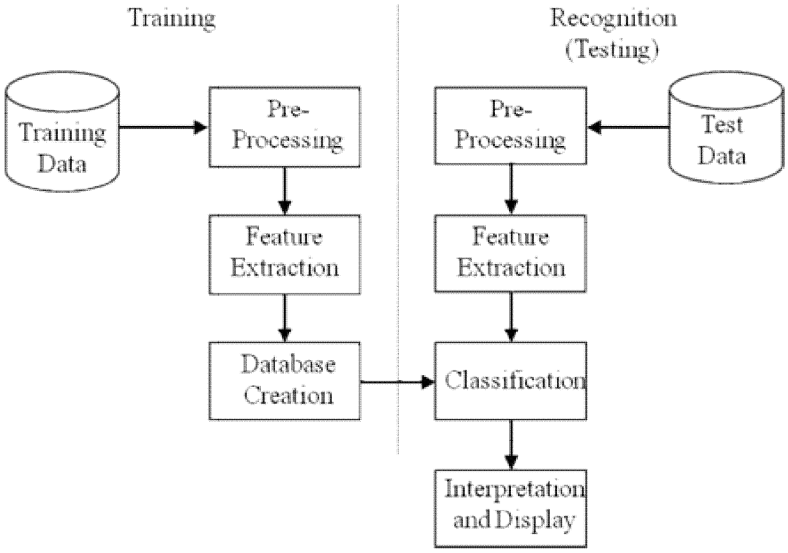

Fig. 1. Block diagram of the proposed method

characters are compared with those in the database. Based on the matched features, the characters are recognized after which the character is interpreted depending on the label assigned previously [20]. After interpretation, the identified text is formatted and displayed in an editable window like notepad.

5. FLOW DIAGRAM

The input document image is read and preprocessed where it is converted from RGB to binary (binarization) and then the background is removed. The input image may contain objects lesser than 15 pixels or irregular or discontinuous

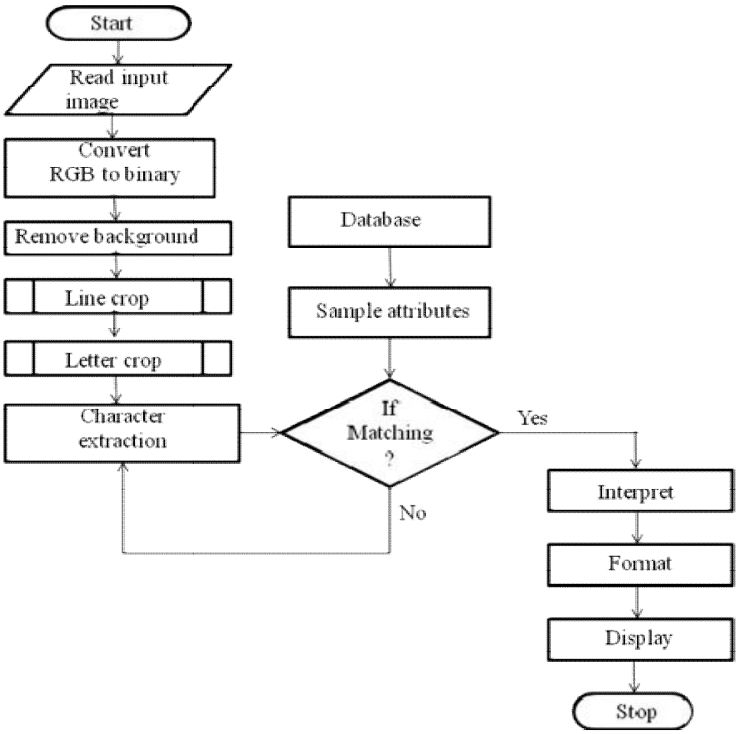

Fig. 2. Flowchart

characters which are considered as noise and are eliminated. The most important reason to remove noise is to remove the extraneous features which otherwise may cause subsequent errors during recognition [21].

The region of interest in the scanned image is extracted by partitioning into multiple segments in the segmentation and is sent for subsequent processing. The relevant information about the characters is extracted by analyzing specific features in the image. Two fundamental functions namely line crop and letter crop are used for character extraction. These features are then interpreted and classified accordingly. Then the extracted individual character is resized to the dimensions of the sample image [22]. Then they are compared with those in the database. Based on the matched features, the identified characters are formatted and displayed in an editable window.

6. RESULTS AND DISCUSSION

In this paper, the scanned/camera-captured document image is the input image. The colour image in RGB format is first converted to a binary image using binarization. Fig. 3.a, Fig. 4.a. and Fig. 5.a. as the input images, the binary images are shown in Fig. 3.b. Fig. 4.b. and Fig. 5.b.

Example 1:

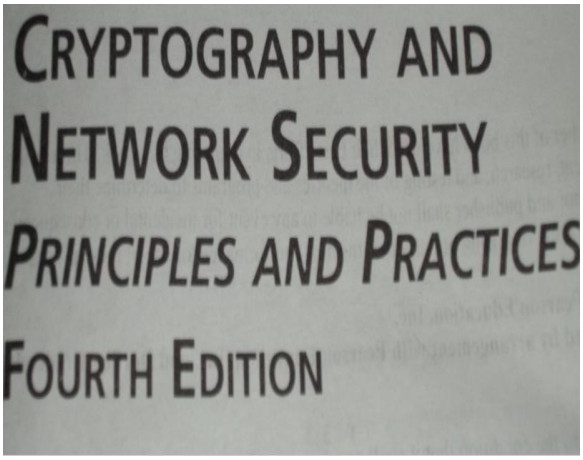

Fig. 3.a. Input image

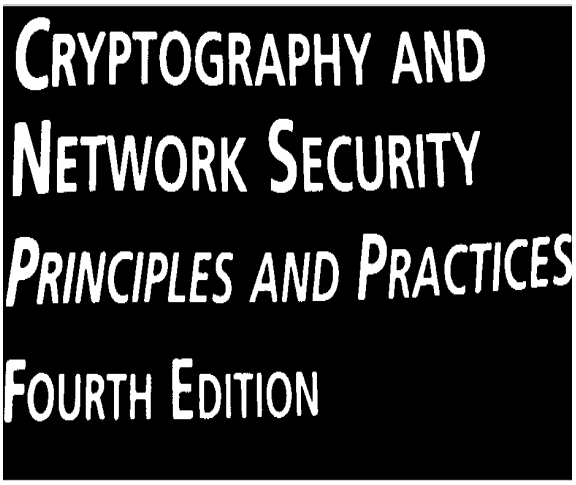

Fig. 3.b. Binary image

Result obtained:

CRYPTOGRAPHY AND
NETORK SECURATY
PWOPLE5 AND PRACT4CES
FOURTH EDSTSON

Output:

CRYPTOGRAPHY AND
NETWORK SECURITY
PRINCIPLES AND PRACTICES
FOURTH EDITION

Example 2:

Fig. 4.a. Input image

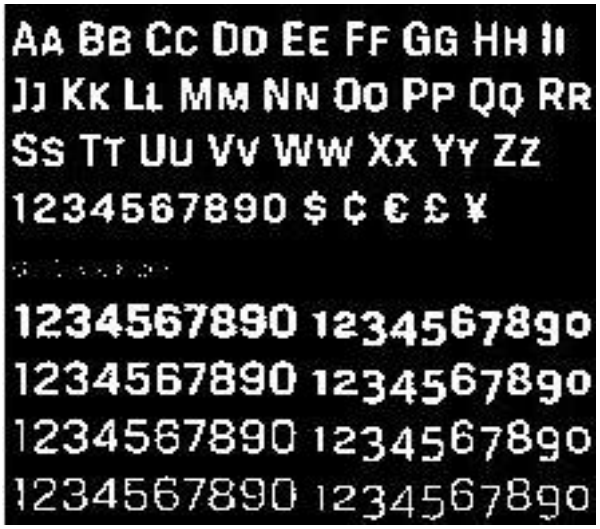

Fig.4.b. Binary image

Result obtained:

AA B9 CC D0 FE FFGG HH 9P
11 KK LL MM NN 00 PP 00 RP
SS TT UU VV WW XX YV ZZ
T234567S90 6 C C C Y
1Z34567890 1234567890
1Z3456789O 1334567890
123456789O 123456789O
1234BB789D Q234BB7B9O

Output:

AA BB CC DD EE FF GG HH II
JJ KK LL MM NN 00 PP QQ RR
SS TT UU VV WW XX YV ZZ
1234567S90 $ ¢ € £ ¥
1234567890 1234567890
1234567890 1334567890
1234567890 1234567890
1234567890 1234567890

Example 2:

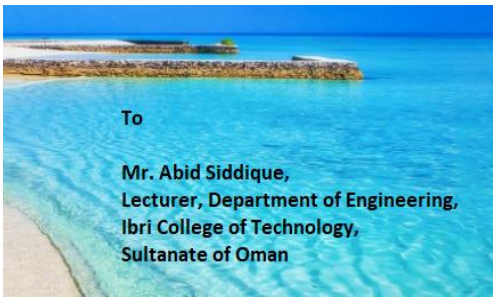

Fig. 5. a. Input Image

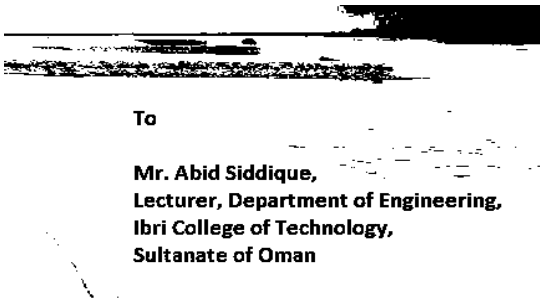

Fig. 5. b. Binary Image

Result obtained:

To,

Mr. Abid Siddique,
Lecturer, Department of Engineering
Ibri College of Technology,
Sultanate of Oman

It can be seen that irregularities in misalignment due to manual errors while scanning are reduced. The output is displayed irrespective of colour, space, font size and intensity in an editable window like Notepad, WordPad where it can even be edited.

7. FUTURE WORK

Direct reading of the text from binary image or colour image is a challenging task because of the complex background or degradations introduced during the scanning of a paper document. In this paper, a simple solution to this problem based on segmentation and feature extraction is presented. This algorithm gives promising results that have been obtained on a number of images in which almost all characters are retrieved. It also gives 90 percent accuracy for all printed characters. In the future, the proposed work can be enhanced by increasing the database to contain alphabets, numbers of any font style which even includes different handwritten styles. However, the one disadvantage of having both

numbers and alphabets in the database is the possibility of misinterpretation of digits having similar features like '0' and 'O', '1' and '*l*' and the like.

New methods for character recognition are still expected to appear, as the computer technology develops and decreasing computational restrictions open up for new approaches. However, the greatest potential seems to lie within the exploitation of existing methods, by mixing methodologies and making more use of context [23]. Generally, there is a potential in using context to a larger extent than what is done today. In addition, combinations of multiple independent feature sets and classifiers, where the weakness of one method is compensated by the strength of another, may improve the recognition of individual characters [24]. The frontiers of research within character recognition have now moved towards the recognition of cursive script that is handwritten connected or calligraphic characters. Promising techniques within this area, deal with the recognition of entire words instead of individual characters.

8. CONCLUSION

This paper focuses on text detection to effectively extract image regions containing text. Our algorithm effectively addresses the three challenging problems in text detection: character category uncertainty, structure pattern variation, and background outlier elimination. We have proposed a structure correlation model to extract discriminative appearance features of characters by local descriptors. Although this technique is having a few shortcomings in cases where a character image is not completely connected, yet it can be used in a variety of applications where the extracted character images are used to train and test a character recognition system like a car number plate recognition system etc. Further, this technique gives future direction for the development of a character extraction technique to extract touching characters like in case of cursive handwriting where characters are touched in a word image. The main contributions of this paper are distinguishing text characters from non-text background outliers, detecting text regions from natural scene images with a complex background, identifying characters from 62 categories, including 10 digits and 26 English letters in upper and lower cases. The method adopted can read the text from scanned document images as well as camera captured images of all major image file formats such as jpg, bmp, gif, tif and png in any font size, shape, intensity, space, color etc. and display it in an editable window like Notepad, WordPad, MS Word with good accuracy.

CONSENT

It is not applicable.

ETHICAL APPROVAL

It is not applicable.

ACKNOWLEDGEMENTS

The authors acknowledge their sincere thanks to the Management of Ibri College of Technology, Oman; for providing infrastructural facilities, support and co-operation to carry out the work successfully in the Institute. The authors particularly thank Mr. Nasser Al Shammakhi, Head of the Department of Engineering, Ibri College of Technology, for his helpful comments and suggestions on a prior version of this review, and the anonymous journal reviewers whose extensive observations certainly improved the quality of this manuscript.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. O'Gorman L. Binarization and multi-thresholding of document images using connectivity, CVGIP: Graphical Models and Image Processing.1994;56.
2. Wahl F, Wong K, Casey R. Block segmentation and text extraction in mixed text/Image documents, Computervision. Graphics and Image Processing. 2001;20: 375-390.
3. Ohya J, Shio A, Akamatsu S. Recognizing characters in scene images, IEEE Trans. Pattern Anal. Machine Intell. 1999;16:215–220.
4. Suen HM, Wang JF. Text string extraction from images of color printed documents, Proc. Inst. Elect. Eng. Vis. Image, Signal Process. 2006;143(4):210–216.
5. Wang L, Pavlidis T. Direct gray-scale extraction of features for character recognition, IEEE Trans. Pattern Anal. MachineIntell. 2003;15:1053–1067.
6. Jagath Samarabandu, Member, IEEE, Xiaoqing Liu. An Edge-based text region

extraction algorithm for indoor mobile robot navigation. International Journal of Signal Processing. 2007;3(4).
7. Hori O, Okazaki A. High quality vectorization based on a generic object model, in Structured Document Image Analysis. Springer-Verlag. 1992;325-339.
8. Suen HM, Wang JF. Text string extraction from images of colour-printed documents, IEEE Proceedings of Vision. Image and Signal Processing. 1996;143:210-216.
9. Sobottka K, Bunke H, Kronenberg H. Identification of text on colored book and journal covers. Document Analysis and Recognition, Bangalore. 1999;57-62.
10. Yuan Q, Tan CL. Text extraction from gray scale document images using edge information. Washington. 2001;302-306.
11. Raju SS, Pati PB, Ramakrishnan AG. Gabor filter based block energy analysis for text extraction from digital document images. Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. 2004;233-243.
12. Shi Z, Setlur S, Govindaraju V. Text extraction from gray scale historical document images using adaptive local connectivity map. Proceedings of the 8th International Conference on Document Analysis and Recognition. 2005;794-798.
13. Qiao YL, Li M, Lu ZM, Sun SH. Gabor filter based text extraction from digital document images. Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, USA. 2006;297-300.
14. Grover S, Arora K, Mitra SK. Text extraction from document images using edge information. IEEE, Gujarat, 2009;1-4.
15. Audithan S, Chandrasekaran RM. Document text extraction from document images using haar discrete wavelet transform. European Journal of Scientific Research. 2009;36:502-512.
16. Hoang TV, Tabbone S. Text extraction from graphical document images using sparse representation. International Workshop on Document Analysis Systems. 2010;143-150.
17. Li Z, Luo J. Resolution enhancement from document images for text extraction. 5th International Conference on Multimedia and Ubiquitous Engineering Loutraki. 2011;251-256.
18. Malakar S, Halder S, Sarker R, Das N, Basu S, Nasipuri M. Text line extraction from handwritten document pages using spiral run length smearing algorithm. International Conference on communications. Devices and Intelligent Systems, Kolkata. 2012;616-619.
19. Rafael C. Gonzales & Richard. E. Woods, Digital Image Processing, 2nd ed. Pearson education; 2001.
20. Rafael C. Gonzales & Richard. E. Woods, Digital Image Processing using MATLAB, 2nd ed. Pearson education; 2001.
21. Anil K. Jain. Fundamentals of digital image processing. Pearson Education, 2001.
22. Yassin MY, Hasan, Lina J. Karam. Morphological text extraction from images. IEEE Transaction on Image Processing. 2000;9(11).
23. Williams PS, Alder MD. Generic texture analysis applied to newspaper segmentation. IEEE International Conference on Neural Networks. 1996;3(6):1664-1669.
24. Zhong Yu, Karu K, Jain AK. Locating text in complex color images. Proceedings of the Third International Conference on Document Analysis and Recognition. 1995;146-149.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history/26071