

Reading and Interpreting Machine Printed Text in Camera Captured Document Images

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 tool box.

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 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 like reading a document and writing it down just as human would, is becoming a reality. The objective of document image analysis is to recognize the text and graphic components in images and extract the intended information as human would. Some tasks are recognizing the text by OCR by determining the skew, columns, text lines and words. 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 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. Character extraction is done in order to speed up the data entry. 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. A system of direct reading of text from documents is introduced here which can be used as an alternative to the existing technology. This paper aims at performing character extraction, recognition and interpretation of character from document images. The result is then exported to spread sheets where it can even be edited.

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.

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

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. Document analysis systems will become increasingly more evident in the form of everyday document systems. 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. METHODOLOGY

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 [8]. Execution is divided into two main stages namely training and testing as shown in figure 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:

2.1 Pre-processing: A scanned document image is considered as input and it is converted from RGB to binary in pre-processing stage, the process is known as binarization and background is removed. The image resulting from the scanning process may contain a certain amount of noise and hence, pre-processing is required.

2.2 Feature extraction: The relevant information about the characters is extracted from the pre-processed image in feature extraction stage and is partitioned into multiple segments during segmentation for subsequent process. Errors in the segmentation process may also result in confusion between text and graphics or between texts and noise.

2.3 Classification: These features are then interpreted and classified accordingly in classification. Structural classification approach is adopted here. Incorrect classification may also be due to poor design of the classifier.

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

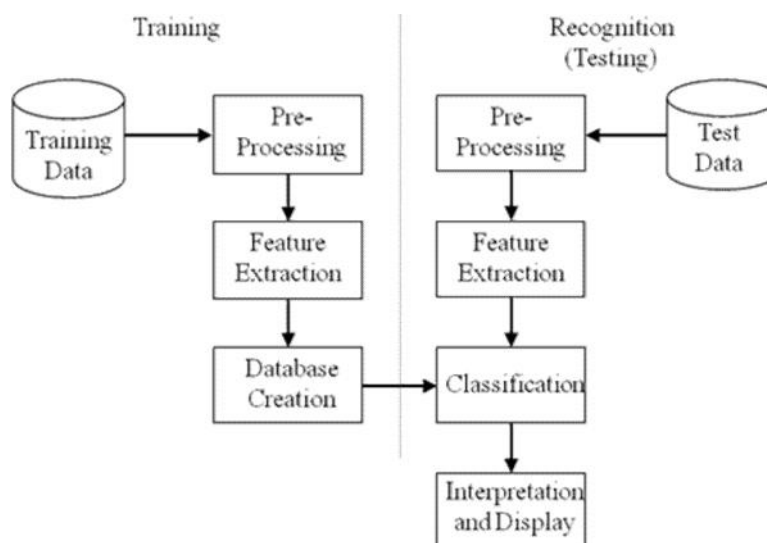


Fig. 1. Block diagram of proposed method

The extracted features are then stored in database for further processing. In identification stage, features of the extracted 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 [9]. After interpretation, the identified text is formatted and displayed in an editable window like notepad.

4. 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 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 [10].

Then 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 [11]. 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.

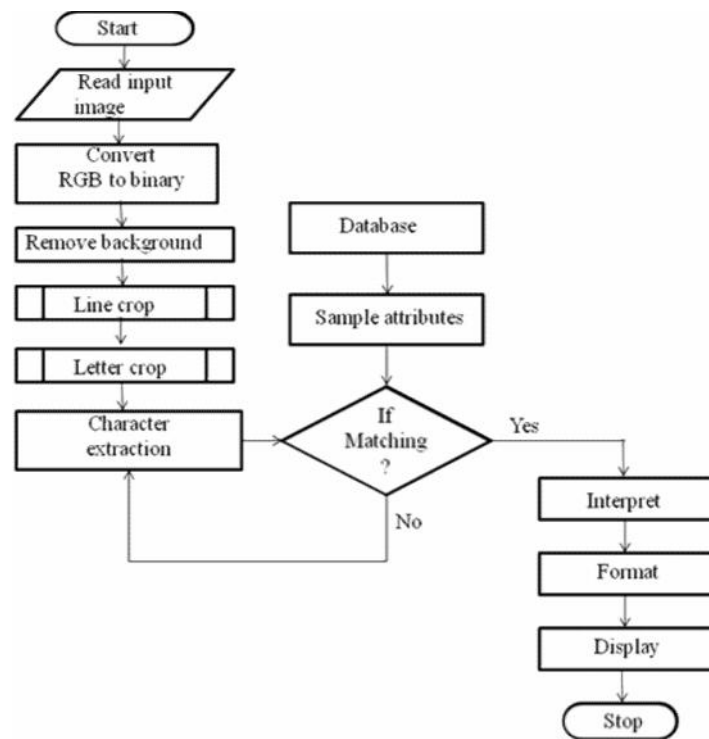


Fig. 2. Flow chart

5. 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. Considering Fig. 3.a. and Fig. 4.a. as the input image, the binary images are shown in Fig. 3.b. and Fig. 4.b.

5.1 Example 1:

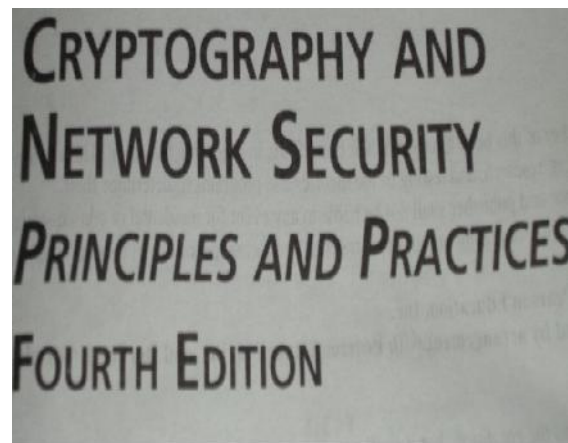


Fig. 3.a. Input image

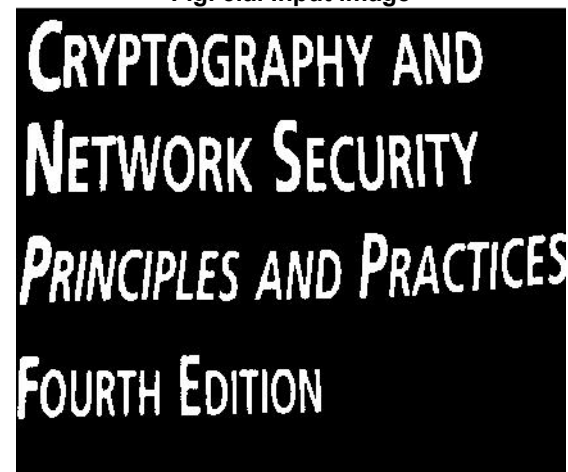


Fig. 3.b. Binary image

Result obtained:

CRYPTOGRAPHY AND
NETORK SECURATY
PWOPLE5 AND PRACT4CES
FOURTH EDSTSON

Output:

CRYPTOGRAPHY AND
NETWORK SECURITY
PRINCIPLES AND PRACTICES
FOURTH EDITION

5.2 Example 2:



Fig. 4.a. Input image

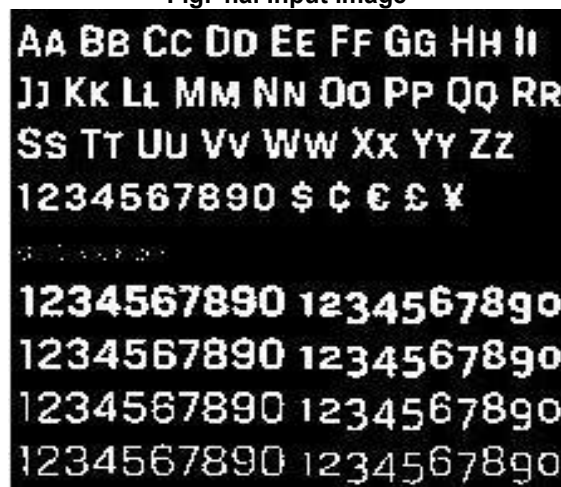


Fig.4.b. Binary image

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
1Z34567890 1334567890
1234567890 1234567890
1234BB789D Q234BB7B90

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

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.

6. FUTURE WORK

Direct reading of 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. There might for instance be a potential in performing character recognition directly on grey level images. However, the greatest potential seems to lie within the exploitation of existing methods, by mixing methodologies and making more use of context [12]. 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 [13]. 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.

7. CONCLUSION

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 (NOT APPLICABLE)

ETHICAL APPROVAL (NOT APPLICABLE)

REFERENCES

1. L. O'Gorman, Binarization and Multi-Thresholding of Document Images using Connectivity, CVGIP: Graphical Models and Image Processing, 1994;56
2. F. Wahl, K. Wong and R. Casey, Block segmentation and text extraction in mixed text/Image documents, Computervision, Graphics and image processing, 2001;20:375-390.
3. J. Ohya. A.Shio, and S.Akamatsu, Recognizing characters in scene images, IEEE Trans. Pattern Anal. Machine Intell., 1999;16:215-220.
4. H. M. Suen and J. F.Wang, Text string extraction from images of color printed documents, Proc. Inst. Elect. Eng. Vis., Image, Signal Process., 2006;143(4):210-216.
5. L. Wang and T. Pavlidis, Direct gray-scale extraction of features for character recognition, IEEE Trans. Pattern Anal. MachineIntell., 2003;15:1053-1067.
6. Jagath Samarabandu, Member, IEEE, and Xiaoqing Liu, An Edge-based Text Region Extraction Algorithm for Indoor Mobile Robot Navigation, International Journal of Signal Processing, 2007;3(4).

269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288

7. O. Hori, A. Okazaki, High quality vectorization based on a generic object model, in Structured Document Image Analysis, Springer-Verlag, 1992;325-339.
8. Rafael C. Gonzales & Richard. E. Woods, Digital Image Processing, 2nd ed. Pearson education; 2001
9. Rafael C. Gonzales & Richard. E. Woods, Digital Image Processing using MATLAB, 2nd ed. Pearson education; 2001.
10. Anil K. Jain, Fundamentals of Digital Image Processing, Pearson Education, 2001.
11. Yassin M. Y. Hasan and Lina J. Karam 2000, Morphological Text Extraction from Images, IEEE Transaction on Image Processing, 2000;9(11).
12. Williams. P. S. and Alder, M. D, Generic texture analysis applied to newspaper segmentation, IEEE International Conference on Neural Networks, 1996;3(6):1664-1669.
13. Zhong, Yu., Karu, K., and Jain, A.K, Locating text in complex color images. Proceedings of the Third International Conference on Document Analysis and Recognition, 1995;146-149.