Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 3222 – 3227

Procedia Engineering

www.elsevier.com/locate/procedia

ICMOC-2012

Computer Assisted Printed Character recognition in document based images

Sushila Aghav,Prof.S.S.Paygude

PG Student ,ME.Computer,MIT,Pune,411038
H.O.D.Computer Engineering,MIT,Pune,411038

Abstract

Printed character recognition is important in the context of document image analysis. Document image analysis analyzes the document images to extract the text and graphics information from image. Text extraction is followed by the character recognition .This paper reviews various phases of machine printed character recognition in document image.

Keywords: Character recogngition;Document image analysis;Template matching

1. Introduction

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 [1].

1. Textual processing deals with the text components of a document image.
2. Graphics Processing

Textual processing tasks are:

1. Determining the skew (any tilt at which the document may have been scanned into the computer),
2. Finding columns, paragraphs, text lines, and words
3. Recognizing the text (and possibly its attributes such as size, font etc.)

* Corresponding author. Tel.:9766464410
E-mail address: sushila_aghav@yahoo.com

1877-7058 © 2012 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.06.373

1.1. Document Based Image Processing

Document Image Text present in images contains useful information for automatic annotation, indexing, and structuring of images [1][2]. Extraction of this information involves detection, localization, tracking, extraction, enhancement, and recognition of the text from a given image. However, variations of text due to differences in size, style, orientation, and alignment, as well as low image contrast and complex background make the problem of automatic text extraction extremely challenging. While comprehensive surveys of related problems such as face detection, document analysis, and image indexing can be found, the problem of text information extraction is not well surveyed. A large number of techniques have been proposed to address this problem.

Document Image Processing phases can be described as follows:

1. Color Image to Gray scale conversion
2. Grayscale to binary conversion
3. Image Binarization (Foreground, Background separation)
4. Text and Non Text Region Separation
5. Text Extraction
6. Character Segmentation
7. Character recognition

The document image captured by camera or scanned image first converted in grayscale format to minimize the processing complexity. Image is then binarized using Image Binarization techniques. Binarization process separates the background from foreground i.e it separate informative region from non informative part.
The informative regions are processed further to separate text part from non text part. Text extraction extracts text from text region. Extracted text further is input to character segmentation phase wherein the each character is separated and extracted. Character recognition recognizes the extracted character by comparing it with the provided character template. Details of the character recognition phases are explained in following section.

2. Machine Printed Character Recognition

Machine Printed Character Recognition carried out in following phases:

1. Template Character Learning
2. Individual Character Extraction
3. Character Feature Extraction
4. Comparison of features of learned and extracted character

2.1. Template CharcterLearning

Character template is represented by an Image [3]. For each character, multiple images are provided as a input to this phase. Character template Images are learned to extract features of the character.
Generating the learned set is quite simple. It requires that an image file with the desired characters in the desired font be created, and a text file representing the characters in this image file. If a character such as pi, has a multicharacter translation, delimiter should be placed around the translation.

Once the learned set has been read in from the image file and its properties recognized, it can be written out to a "learn" file. This file stores the properties of the learned characters in abbreviated form, eliminating the need for retaining the images of the learned characters, and can be read in very quickly.

a	a	*a*	a	A
b	b	*b*	b	B
C	C	*C*	C	C
D	D	*D*	D	D
E	E	*E*	E	E
F	F	*F*	F	F
G	G	*G*	G	G
H	H	*H*	H	H
1	1	*1*	1	1
5	5	*5*	5	5

Table 1: Sample Character Templates

2.2. Individual Character Extraction

Character extraction divided into two phases.

1. Text line segmentation.
2. Detection of Connected Component[5]

2.2.1. Text Line Segmentation

Various Text line segmentation techniques are [1]

- Projection-based
- Hough Transform,
- Smearing methods,
- Grouping methods,
- Active Contour methods,
- Graph-based methods.

The projection-based approaches are making use of the structural characteristics of the documents. They are top-down techniques, simple and easy in implementation. Hough Transform is also a popular methodology in the area of text line segmentation .It describes parametric geometric shapes and identifies geometric locations that suggest the existence of the sought shape. Serious drawback of this method is the computational complexity. The smearing methodology is a bottom-up technique. It is the process of converting a set of background pixels located between foreground pixels into foreground pixels whether their amount is less than a certain threshold. Smearing methods strengthen by local techniques, solve

specific problems and overlapping touched connected component. Moreover, these methods work successfully with documents that contain characters of variable height.

The grouping methods [1] are also bottom-up. From the lower level, the pixel, starts a process of grouping according to specific constrains designed to result to a layer of text lines. The process is relatively easy in the case of printed documents, but it may be proved to be difficult and problematic in manuscripts.

The Active Contour methods use the difference between the foreground and the background through characteristics such as brightness or color that occurs at the border contours of the object. The edge is a curved line from which derive all the properties and characteristics that describe the specific category of shapes, in our case text lines.

The representation of document images by graphs is an important tool of the line segmentation procedure. The graph is constructed as vertices of pixel or more complex connected components. The vertices are normally associated with weighted edges that depict distances between connected components. After the modeling of the document image, the treatment method can be chosen.

2.2.2. Detection of Connected Component

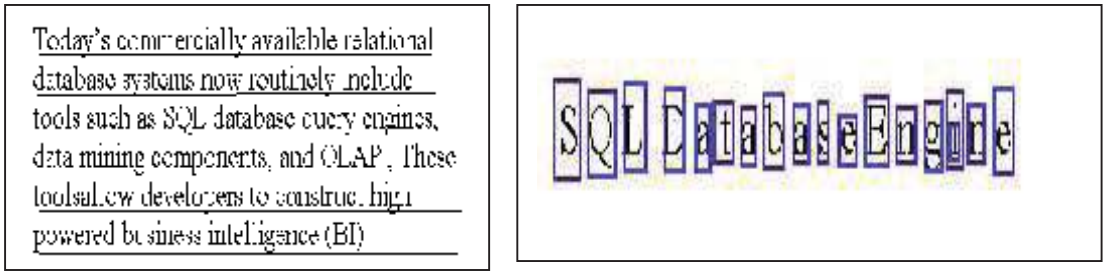

Fig. 1. (a)Text Line Segmentation (b) Connected Component Detection

To extract the connected components from each line, starting at the upper right corner of each line, removes touching intervals of black pixels from the image until nothing more connected can be found. The extraction routine then looks upward and downward to see if there are possible "extra parts", such as the dot on an 'i', hanging directly above or below the component. Character Feature Extraction

Extracted Features of each character will be stored in a vector. For this the character can be divided into equal sized region and intensity values of the pixel in particular region is considered ad feature.

2.3. Comparison of features of learned and extracted character

Template Features are compared with the extracted character feature [7]. For comparison, the correlation analysis is done on feature vectors on template and character image.

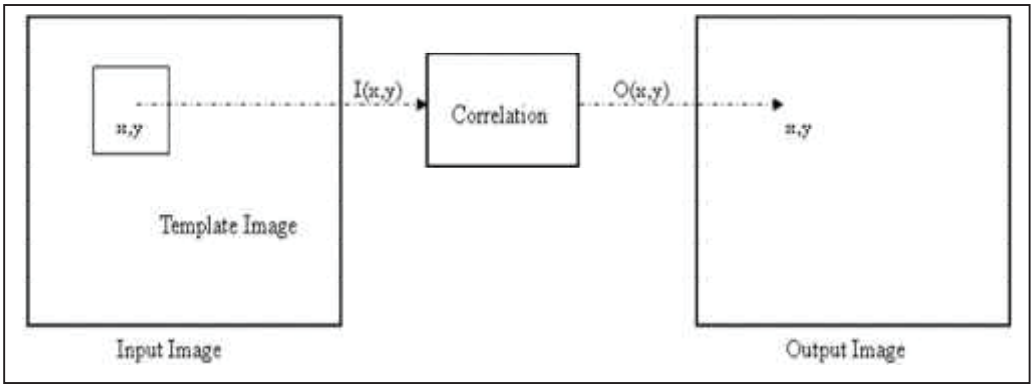

Fig. 2. (a) Correlation Analysis

Co relation Function is shown in *(I)*

$$C = \sum_{x \in N, y \in N} |t(x,y) - t'(x,y)| \qquad (I)$$

Correlation analysis finds out the similarity between the feature vectors of two images. If the similarity coefficient is between 0 and 1.

If the result of correlation analysis is more that 0.5 the match is good enough to recognize the character. And character said to be recognized.

3. Conclusion

In this paper we have reviewed the machine printed character recognition phases. There are various techniques which can be applied at each recognition phase to get correct and efficient recognition results.

Text line segmentation, character segmentation can be implemented in variety of ways, out of those Histogram projection profile found an easy technique for text segmentation. For character extraction the connected component found to be efficient technique. The result of the character recognition can be improved by extending feature vector size and adding more character features.

References

[1] Ergina Kavallieratou,Fotis Daskas, Text Line Detection and Segmentation: Uneven Skew Angles and Hill-and-Dale Writing, *Journal of Universal Computer Science, vol. 17, no. 1 (2010), 16-29.*

[2] Lawrence O'Gorman,Rangachar Kasturi ,Document Image Analysis, ISBN 0-8186-7802-X,Library of Congress Number 97-17283 .

[3]Jonghyun Park, Toan Nguyen Dinh, and Gueesang Lee,,Binarization of Text Region based on Fuzzy Clustering and Histogram Distribution in Signboards, World Academy of Science, Engineering and Technology 43 2008

[4] Keechul Jung, Kwang In Kim, Anil K. Jain, Text Information Extraction in Images and Video: A Survey

[5]J.Sushma, M.Padmaja, Text Detection in Color Images, 978-1-4244-4711-4/09©2009 IEEE,IAME 2009

[6]George Nagy ,Rensselaer Polytechnic Institute, State of Art of Document Image Processing,2008

[7]Ayatullah Faruk Mollah1, Nabamita Majumder, Subhadip Basu and Mita Nasipuri, Design of an Optical Character Recognition System for Camera-based Handheld Devices, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011

[8]Jiang Gao, Jie Yang,"An Adaptive Algorithm for Text Detection from Natural Scenes",2010

.