ijecs
open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume - 3 Issue - 8 August, 2014 Page No. 7733-7737

Character Recognition from Born Digital Images using Correlation

Soumya soman
Department of Computer Science
Govt Engineering College ,Idukki

***Abstract*—Character recognition can be done on various image datasets. In this method Binarization and edge detection are separately carried out on the three colour planes of the image. From the binarized image Connected components (CC's) are obtained and thresholded based on their area and aspect ratio. CC's which contain sufficient edge pixels are retained. Also the text components are represented as nodes of a graph. The centroids of the individual CC's are represented as the nodes of the graph. Long edges are broken from the minimum spanning tree of the graph. Pairwise height ratio is also used to remove likely non-text components. A new minimum spanning tree is created from the remaining nodes. Horizontal grouping is performed on the CC's to generate bounding boxes of text strings. Overlapping bounding boxes are removed using an overlap area threshold. Non-overlapping and minimally overlapping bounding boxes are used for text segmentation. Vertical splitting is applied to generate bounding boxes at the word level . After the segmentation and Localization text is character rs are recognized by using Correlation.**

***Keywords*- Character Recognition ; Binarization ; Minimum spanning tree; Text segmentation; Text localization ;Text Correlation**

I.INTRODUCTION

People have always tried to develop machines which could do the work of a human being. The reason is obvious since for most of history, man has been very successful in using the machines developed to reduce the amount of physical labor needed to do many tasks. With the advent of the computer, it became a possibility that machines could also reduce the amount of mental labor needed for many tasks. Over the past fifty or so years, with the development of computers ranging from ones capable of becoming the world chess champion to ones capable of understanding speech, it has come to seem as though there is no human mental faculty which is beyond the ability of machines. Today, many researchers have developed algorithms to recognize printed as well as handwritten characters. But the problem of interchanging data between human beings and computing machines is a challenging one. In reality, it is very difficult to achieve 100% accuracy. Even humans too will make mistakes when come to pattern recognition. The accurate recognition of typewritten text is now considered largely a solved problem in applications where clear imaging is available such as scanning of printed documents. Typical accuracy rates on these exceed 99%; total accuracy can only be achieved by human review. Other areas including recognition of hand printing, cursive handwriting, and printed text in other scripts especially those with a very large number of characters are still the subject of active research. It is a part of pattern recognition that usually deals with the realization of the written scripts or printed material into digital form.

Text embedded in landmarks, milestones, display boards, and building names acts as one of the visual aids. Text, if present in an image, contributes to tagging information. An image with text superimposed by a software is known as born-digital image. Born-digital images are used in web pages and e-mail as name, logo or ads. The resolution of the text present in the image and anti-aliasing of text and the background form the major differences between scenic and born-digital images. Unlike born-digital images, a large amount of variation in the background may be present in scenic images due to illumination changes. In this paper, we propose an algorithm for Character Recognition from born digital images. Here we use the concept of OCR for recognizing characters from the image.

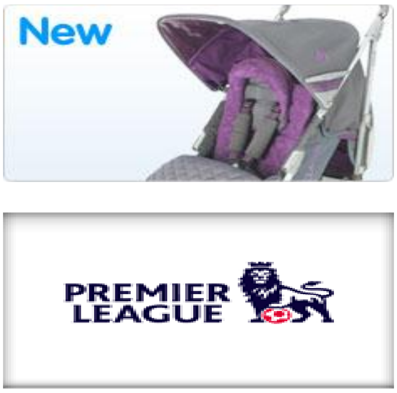


Figure 1 : Sample images from Born-Digtial image competition dataset

The main advantage of storing these written texts in digital form is that, it requires less space for storage and can be maintained for further references without referring to the actual script again and again.

A. Image and Image Processing

Image is a two-dimensional function *f(x,y)*, where x and y are spatial coordinates and the amplitude f at any pair of coordinates *(x,y)* is called the intensity or gray level. When x, y, and f are discrete quantities the image is digital. 'f' can be a vector and can represent a color image, e.g. using the RGB model, or in general a multispectral image.

The digital image can be represented in coordinate convention with M rows and N columns. In general, the gray-level of pixels in an image is represented by a matrix with 8-bit integer values.

Image Processing is all about improvement of pictorial information for human interpretation and processing of image data for storage, transmission and representation for autonomous machine perception. Processing of image data enables long distance communication, storage of processed data and also for application which require extraction of minute details from a picture. Digital image processing concerns with the transformation of an image to a digital format and its processing is done by a computer or by a dedicated hardware. Both input and output are of digital in nature. Some processing techniques tend to provide the output other than an image which may be the attributes extracted from the image, such processing is called digital image analysis. Digital image analysis concerns the description and recognition of the image contents where the input is a digital image; the output is a symbolic description or image attributes. Digital Image Analysis includes processes like morphological processing, segmentation, representation & description and object recognition (sometimes called as pattern recognition). Pattern recognition is the act of taking in raw data and performing an action based on the category of the pattern. Pattern recognition aims to classify data (patterns) based on the information extracted from the patterns. The classification is usually based on the availability of a set of patterns that have already been classified or described. One such pattern is Character.

The main idea behind character recognition is to extract all the details and features of a character, and to compare it with a standard template. Thus it is really necessary to segment these characters before proceeding with the recognition techniques. To achieve this, the printed material is stripped into lines, and then into individual words. These words are further segmented into characters.

B. Characters - An overview

Characters in existence are either printed or handwritten. The major features of *printed characters* are that they have fixed font size and are spaced uniformly and they do not connect with its other neighboring characters. Whereas handwritten characters may vary in size and also the spacing between the characters could be non-uniform.In born digital images the characters are in printed format.

Processing of printed characters is much easier than that of handwritten characters. By knowing the spaces between each character in printed format, it is easy to segment the characters. For handwritten characters, connected component analysis has to be applied, so that all the characters can be extracted efficiently. Although there are 26 characters in English language, it is observed that both uppercase and lowercase letters are utilized during the construction of a sentence. Thus, it is necessary to design a system which is capable of recognizing a total of 62 elements (26 lowercase characters + 26 uppercase letters + 10 numerical).

II. RELATED WORRK

In 1929 Gustav Tauschek obtained a patent on OCR in Germany, followed by Paul W. Handel who obtained a US patent on OCR in USA in 1933. In 1935 Tauschek was also granted a US patent on his method. Tauschek's machine was a mechanical device that used templates and a photo detector.

In 1949 RCA engineers worked on the first primitive computer-type OCR to help blind people for the US Veterans Administration, but instead of converting the printed characters to machine language, their device converted it to machine language and then spoke the letters. It proved far too expensive and was not pursued after testing

In 1950, David H. Shepard, a cryptanalyst at the Armed Forces Security Agency in the United States, addressed the problem of converting printed messages into machine language for computer processing and built a machine to do this, reported in the Washington Daily News on 27 April 1951 and in the New York Times on 26 December 1953 after his was issued. Shepard then founded Intelligent Machines Research Corporation (IMR), which went on to deliver the world's first several OCR systems used in commercial operation.

In 1955, the first commercial system was installed at the Reader's Digest. The second system was sold to the Standard Oil Company for reading credit card imprints for billing purposes. Other systems sold by IMR during the late 1950s included a bill stub reader to the Ohio Bell Telephone Company and a page scanner to the United States Air Force for reading and transmitting by teletype typewritten messages. IBM and others were later licensed on Shepard's OCR patents.

III.ALGORITHM

The algorithm for character recognition from born digital images is performed by using segmentation,, pruning of connected components, Horizontal grouping, minimal spanning tree etc. The proposed algorithm is split into sections for ease of description. The processing involved in some of the sections are required to be performed repeatedly on individual colour planes.

1.SEGMENTATION

Initially k-means clustering was used to form clusters. Due to low resolution characters, small width characters may be merged into the background cluster .Therefore, a thresholding algorithm was used in segmentation task.By using ‘Otsu’ global thresholding technique [3] ,each of the R, G and B colour planes are separately segmented. In this method, the threshold calculation is posed as an optimization problem and it is maximized. This thresholding scheme is used in binarization of document images, and works well for images of good quality. The connected components (CC’s) in the binarized image are labeled. For possible text components ,the complement version of binarized image is also considered, to account for the presence of inverse text. Edges are detected for individual colour planes using Canny’s [4] edge detection algorithm. The edge pixels are termed as edgels. These edgels are used to select the CC’s from the binarized image.

2.PRUNING OF CONNECTED COMPONENTS

This part of algorithm is carried out on each of the binarized planes and their complements. To counter the polarity of the text existing in an image complement form is used.Tto remove possible non-text components ,the area, bounding box and pixel list are obtained for each CC and are thresholded . There may have a lot of stray pixels resulting from thresholding. For removing the stray pixels the CC's with more than 5 pixels were only retained. Aspect ratio is the ratio of height to width. Aspect ratio is calculated for each CC using the dimensions of the bounding box of CC. The range of aspect ratio is fixed between 0.01 to 20. The CC's which have pixels on the image boundary are removed; this ensures that broken characters do not appear in the next step of the algorithm. In the case of low resolution text characters, there will be a displacement one pixel in the Canny edge operation. Hence, the CC's are morphologically thickened. The thickened CC's are placed on the edge map of the colour plane and the number of edgels present within the thickened CC is counted. Non-text components do not have edgels within it. The components with edgel count less than 6 are removed. The training dataset had only English characters. The maximum Euler number for English characters is 2. The CC's whose Euler number is greater than 2 are also removed. The CC's preserved in the binarized plane and its complemented version are morphologically thinned and clubbed together to form a single thinned image plane.

3.MINIMUM SPANNING TREE

Minimum spanning tree algorithm is performed separately on each of the thinned images obtained from the three colour channels. Also the connected components are labeled in each of these images. The centroid of these CC's are used as the nodes of a graph. We use the fact that ,in a text string, the characters are closer when compared to non-text components.By using tree algorithm proposed by Prim [5] a minimum spanning tree is generated for the graph.This algorithm uses shortest distance between the nodes and bridges the nodes to generate spanning tree for the graph. By using a length threshold 2.5 times the mean value of the edges present in the spanning tree ,the isolated nodes with long edges are broken. These isolated nodes are presumed to be non-text components appearing in the thinned image. Also we have a text string has characters of similar height. The preserved nodes of spanning tree are subjected to mutual height ratio test. The height of each CC is compared with other CC's. If more than two CC's have height ratio in the range 0.5 to 2, then that CC is retained. The retained CC's are converted into nodes and another pass of minimum spanning tree is carried out. The mean value of the edges from minimum spanning tree is reduced if non-text CC's have been removed. Some of the non-text components with heights similar to that of text components may not be removed by the above check. These non-text components are usually far away from the string of characters. Second pass of minimum spanning tree is to ensure the removal of those non-text components also. The result of this section is used to retrieve the actual connected components from the binarized image. The connected components from all the three binarized images are grouped into a single plane for text segmentation and localization.

4.HORIZONTAL GROUPING OF TEXT COMPONENTS

We horizontally group appropriate CC's into words since most of the training images had only horizontal text. Based on the earliest occurring top row of the CC, the CC's from the combined image plane are sorted in ascending order, The bounding box and centroid value of each CC are obtained.Grouping is performed on the CC's with centroids falling within the vertical range of the bounding box of a candidate CC. Until all CC's have been grouped ,this process is repeated. We have the fact that non-text components which pass through the minimum spanning tree module had large overlap with other components. Hence, the bounding box values of grouped CC's are tested for overlap condition, and those with high overlap are removed. Also,the bounding box should be removed if the area of its overlap with another BB is more than 20 percent of its own area. Also the bounding boxes containing a single CC are also removed, since we are interested only in bounding boxes containing words. The CC's from the selected bounding boxes are preserved as segmented text at the level of pixels. For the presence of multiple words ,bounding boxes containing more than 4 CC's are tested. If the maximum value of gap between two neighboring CC's is higher than the mean value of the intercomponent gaps by more than 3 pixels, then a threshold of (max value − 2) pixels is used to split the bounding box into two. If the height of the bounding box of the horizontally grouped CC's is less than 11 pixels, a fixed threshold of 3 pixels is used to split the bounding box. This splitting into individual words is termed as vertical splitting. After this, the bounding box values are calculated and the text localization details are created at the output stage. Vertical splitting block outputs the final bounding boxes. The bounding box list provides the text localization information. Selected CC's form the segmented text at the pixel level.

5.CHARACTER RECOGNITION USING CORRELATION

Correlation is a signal-matching technique. It is an important component in digital communication system. It is often used in signal processing for analyzing functions or series of values, such as time domain signals. A correlation is useful because it can indicate a predictive relationship that can be exploited in practice.

a) Correlation

In signal processing correlation can be defined as the technique which provides the relation between any two signals under consideration. The degree of linear relationship between two variables can be represented in terms of a Venn diagram. Perfectly overlapping circles would indicate a correlation of 1, and non-overlapping circles would represent a correlation of 0. For example questions such as "Is X related to Y?", "Does X predict Y?", and "Does X account for Y?" indicate that there is a need for measuring and better understanding of the relationship between two variables. The correlation between any two variables 'A' and 'B' can be denoted by *"RAB"* .Relationship refers to the similarities present in the two signals. The strength of the relation will always be in the range of 0 and 1. The two signals can be said to be completely correlated if the strength of their relationship is 1 and are completely non-correlated if the strength of the relationship is 0.

In character recognition, each character can be considered as an image and hence 2D correlation can be implemented for character recognition. Before starting the recognition process, the result of the horizontal grouping of text to go through some preliminary stages where the image can actually be processed so that it can be used to recognize the characters present in it.

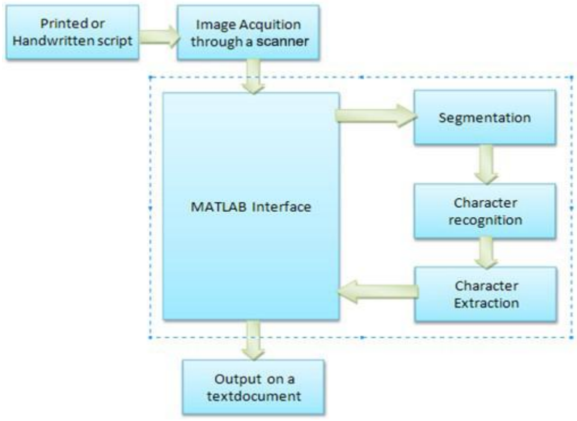
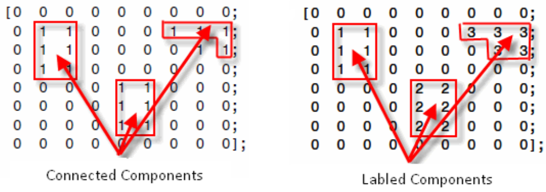


Figure 2: Block diagram for implementing recognition process

The characters are extracted through a process called connected component analysis. First the image divided into two regions. They are black and white region. Using 8-connectivity the characters are labeled. Using these labels, the connected components (characters) are extracted. The extracted characters are then resized to *35 X 25*.



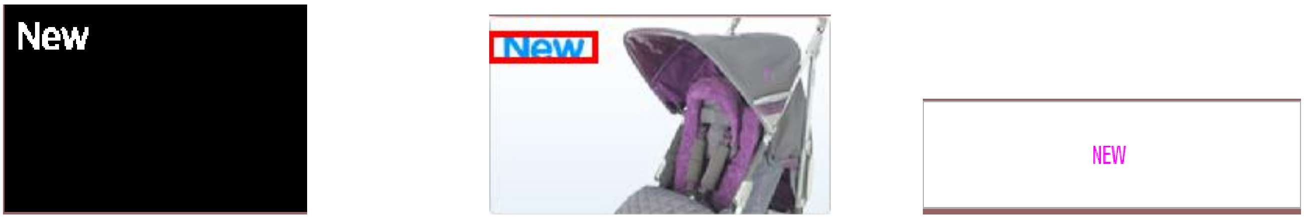
```
[0  0  0  0  0  0  0  0  0;
 0  1  1  0  0  0  1  1  1;
 0  1  1  0  0  0  0  1  1;
 0  1  1  0  0  0  0  0  0;
 0  0  0  0  1  1  0  0  0;
 0  0  0  0  1  1  0  0  0;
 0  0  0  0  1  1  0  0  0;
 0  0  0  0  0  0  0  0  0];
```

Connected Components

```
[0  0  0  0  0  0  0  0  0;
 0  1  1  0  0  0  3  3  3;
 0  1  1  0  0  0  0  3  3;
 0  1  1  0  0  0  0  0  0;
 0  0  0  0  2  2  0  0  0;
 0  0  0  0  2  2  0  0  0;
 0  0  0  0  2  2  0  0  0;
 0  0  0  0  0  0  0  0  0];
```

Labled Components

Figure 3: (a) connected components; (b) labeling of the connected components



(a) Text Segmentation (b) Text Localization (c) Character Recognition

Figure 4: Result of the proposed algorithm for the first born-digital image in Fig. 1

A connected component in a binary image is a set of pixels that form a connected group. For example, the binary image below has three connected components (figure 3(a)).Connected component labeling is the process of identifying the connected components in an image and assigning each one a unique label (figure 3(b)). The matrix (figure 3(b)) is called a *label matrix*. For visualizing connected components, it is useful to construct a label matrix.

b) Recognition

In the recognition process, each character extracted is correlated with each and every other character present in the database. The database is a predefined set of characters for the fonts Times new roman, Tahoma and Verdana. All the characters in the database are resized to *35 X 25*.

By knowing the maximum correlated value, from the database, the character is identified. Finally, the recognized characters are made to display on a box.That is the recognized outputs for the segmented images.

The algorithms used in this paper take advantage of the fact that the order, speed, and direction of individual lines segments at input are known. Also, the user can be retrained to use only specific letter shapes. These methods cannot be used in software that scans paper documents, so accurate recognition of character from low contrast images is still largely an open problem. Accuracy rates of 80% to 90% on neat, clean low contrasted characters can be achieved, but that accuracy rate still translates to dozens of errors per page, making the technology useful only in very limited applications.

IV. CONCLUSION

Thus we have proposed a new, effective method for character recognition using correlation from born-digital images which is easy to implement. Since this algorithm is based on simple correlation with the database, the time of evaluation is very less.

Also the database which was partitioned based on the areas of the characters made it more efficient. Thus, this algorithm provides an overall performance in both speed and accuracy.In this method,space between characters is achieved through trial and error method. As a future work we Would like to add neural network to improve the correctness of text extraction.

REFERENCES

[1] IAPR TC11 Reading Systems-Datasets List, http://www.iaprtc11/ mediawiki/index.php/Datasets

[2] D. Karatzas, S. Robles Mestre, J. Mas, F. Nourbakhsh and P. Pratim Roy," ICDAR 2011 Robust Reading Competition - Challenge 1: Reading Text in Born-Digital Images (Web and Email)", In Proc. 11th International Conference of Document Analysis and Recognition, 2011 , September 2011, http://www.cv.uab.es/icdar2011competition/

[3] N. Otsu, "A Thresholding Selection Method from Gray-level Histogram", IEEE Transanctions on Systems, Man and Cybernetics,

vol. 9, pp. 62-66, March 1979.

[4] J. Canny, "A Computational Approach to Edge Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pp. 679-698, November 1986.

[5] R. Prim, "Shortest connection networks and some generalizations", Bell System Technical Journal, 38:1389-1401, 1957.

[6] C. Wolf and J. M. Jolin," Object count/area graphs for the evaluation of object detection and segmentation algorithms", International Journal of Document Analysis, 8(4), pp.280-296,
2006.

[7] Deepak Kumar and A G Ramakrishnan, "OTCYMIST: Otsu-Canny Minimal Spanning Tree for Born-Digital Images", IAPR International Workshop on Document Analysis Systems 2012

[8] Negi, C. Bhagvati and B. Krishna, "An OCR system for Telugu", in the Proceedings of the
Sixth International Conference on Document Processing, pp.1110-1114, 2001 .

[9] Dr.-Ing. Igor Tchouchenkov, Prof. Dr.-Ing. Heinz Wörn, "Optical Character Recognition
Using Optimisation Algorithms", Proceedings of the 9th International Workshop on Computer
Science and Information Technologies CSIT'2007, Ufa, Russia, 2007.

[10] Sonka, Halvac, Boyle, "Digital image processing and computer vision", first Indian reprint 2008, page 345-349.

[11] Michael Hogan, John W Shipman, "OCR (Optical Character Recognition): Converting paper
documents to text", thesis submitted to New Mexico Tech Computer Center, 01-02-2008

[12] Rafeal C.Gonzalez, Richard E.Woods, "Digital Image Processing", third edition 2009.