Character Reorganization and Categorization using Hybrid Algorithms by Scanned Documents

Rudihartono Ismail, Ismail Suardi Wekke, A. Dinesh Kumar, R. Pandi Selvam, K. Shankar, Phong Thanh Nguyen

Abstract: Detection and reorganization of text may save a lot of time while reproducing old books text and its chapters. This is really challenging research topic as different books may have different font types and styles. The digital books and eBooks reading habit is increasing day by day and new documents are producing every day. So in order to boost the process the text reorganization using digital image processing techniques can be used. This research work is using hybrid algorithms and morphological algorithms. For sample we have taken an letter pad where the text and images are separated using algorithms. The another objective of this research is to increase the accuracy of recognized text and produce accurate results. This research worked on two different concepts, first is concept of Pixel-level thresholding in image processing and another one is Otsu Method thresholding.

Keywords: Character Recognition Technique, OCR, Segmentation.

I. INTRODUCTION

Enhance an existing image and produce the better output, recognized desired object, text, colour correction etc. is called digital image processing. Every digital image processing research has some common steps. These steps are given as follows:

• Import input image using Image acquisition tools;
• Analyze, resize and manipulate input image;
• Generate output and measure its accuracy

Why Digital Image processing

There are five basic purposes for a digital image processing research. They are:

• Object Visualization
• Increase Sharpness and Restore
• Image retrieval
• Pattern Measure
• Produce Recognized Image.

II. APPLICATIONS OF IMAGE PROCESSING

Day by day new experiments are research is going on in the field of digital image processing. Such researches are useful in many different disciplines covering medicine through remote sensing.

1. Super-high-definition image processing
2. Video processing
3. Color processing
4. Digital cinema
5. High-quality color representation
6. High-resolution display
7. Hybrid techniques
8. Image processing architectures and workstations
9. Image sharpening and restoration
10. Image transmission and coding
11. Multidimensional image processing
12. Pattern recognition
13. Programmable DSPs for video coding
14. Registration techniques
15. Remote sensing
16. Robot vision
Research objective

This research has following main objectives:

- Detect the text from an input document image. To recognize text and produce better results this research compare two thresholding techniques first is Pixel-level thresholding and second is is Otsu Method thresholding.
- The result of research will be shown in the form of tabular comparison.

III. PROPOSED APPROACH

This research used basic morphological functions for characters and header separation using image processing technique in Matlab. And this approach works on two thresholding methods.

This section includes the study of different techniques that uses for developing the algorithm to detect or separate the text from an image.

Character Segmentation

Character segmentation is the process of identifying regions in an image that represent characters. One of the commonly used methods for character segmentation is Projection Analysis, surveyed in (Casey and Lecolinet, 1996), and applied to license plates in (Zhang and Zhang, 2003). This method consists of combining pixel values across one dimension, where peaks & valleys of this projection give insight to the character’s locations. But the problems arises when a bigger amount of background noise and blur are present, or if there is some spacing between characters. More sophisticated segmentation techniques incorporate a form of binarization followed by connected component analysis to distinct characters.

As a result of these way may be good for reading a huge variety of image text domains, but for the specific task, they are not ideal.

Figure 2: Character segmentation is the first step

Figure 3: Taken image from chen and yuille (2004), when images are too blurry, character segmentation fails.

Character Classification

Although methods in character segmentation rarely incorporate machine learning1 , there has been a strong trend in recent years toward its use in character classification. When a single font is present in the data, as is the case with reading Korean license plates in (Yu and Kim, 2000), template matching can be accurate and fast. However, when there is a large amount of noise, obstruction, and various fonts, machine learning can be very powerful.

IV. EXPERIMENTAL RESULT AND ANALYSIS

Analyzing a document layout in image is a very important step in many software and applications related to documents images, such as extracting text or characters using optical character recognition (OCR), configuration the documents flow, and documents arranged according to a particular sequence. The aim of such document-image analysis to convert the matter contained on a digitized image into an equivalent symbolic form. Text in the documents represents information stored in most applications. So in order to extract text it is very necessary to identify text objects within the image, recognize texts, and extract the information in a systemic way.[63]

Data Information

This thesis work is used a image which is shown in figure 4.1. Whole process for detecting the text from images is processed on that image. After that two types of thresholding techniques were used to get the better accuracy.
Methodology
The morphological recognition algorithm and Otsu thresholding methods are used in developing an automated system in MatLab 2013a. In this approach an gray scale image is produced for preprocessing. This image is taken by the camera or scanned image.

Image Acquired
Initially once input image is obtained in image acquisition stage, the pre processing methods can be applied to the image. Such steps may help to produce many visions of the input image. Further such visions can be used for further processing. Figure below shows the acquired image.

Steps in the block diagram:
1) Lette Pad Image Acquisition:
   a) Input document image 
b) Resizing the Document Image
2) Preprocessing and Sharping Image
   a. RGB to Gray Scale Image conversion 
b. Document Image complemented
3) Segmentation of Document
   a. Document Image Binary conversion 
b. Edge Detection 
c. Clustering Method
4) Document Scanned Image Enhancement
   a. Filling Holes on images 
b. Creating Holes Edge Detected Images 
c. Filtration of image using Bewareopen command using High pass filter
5) Object recognition
   a. Recovered text and header using Cluster method.
Image Resize

Resizing of input image is necessary to save the memory. Here image with 400 X 600 resolution is used.

RGB to Gray Scale Conversion

Digital Image processing uses Gray scale image for processing. So that the image is converted into gray scale.

Canny Edge Detection

In order to detect the text it is necessary to identify characters sharp edge. The canny edge detection algorithm fetch useful structured data and provide many useful visions of document edge.
Figure 12: canny edge detected image

**Filling Holes**

Here we will fill the holes which were created on the canny edge detection to extract the text. This is the major step of text extraction. This is the main part of the Morphological operations. Figure of this step is given shown in figure below:

Figure 13: Filling contour holes for text extraction.

**Filteration of Image Document using HighPass Filter**

This method sharps edges of the input image. It increases contrast between characters gap and provides pixels which are greater than 25. Other characters will be assumed as suppressed.

**Separated Text**

The segmentation of document successfully separates header of document and text. The produced output of the document is given below:

Figure 14: Received Text

Figure 15: Received Header or image

4.1 Result

Same method is applied on the reference image using two different thresholding methods. One is pixel level and second is Otsu thresholding. A graph is shown in figure which shows the execution time for both thresholding techquines.

Figure 16: Graph Between two Thresholding Method

Its indicating that Otsu method is 1.32 second faster than the pixel level thresholding method. Output in % = (pixel level thresholding – Otsu Method thresholding ) /Otsu method thresholding in % = 33.426%

As the result showing we can concluded that Otsu method is faster than pixel level thresholding more than 30.62%.

V. CONCLUSION

Recognition and rearrangement of content may spare a ton of time while imitating old books content and its parts. This is truly testing exploration point as various books may have diverse textual style types and styles. The computerized books and eBooks perusing propensity is expanding step by step and new records are delivering each day.
So as to support the procedure the content revamping utilizing computerized picture preparing methods can be utilized. This examination work is utilizing half breed calculations and morphological calculations. For test we have taken a letter cushion where the content and pictures are isolated utilizing calculations. The target of this exploration is to expand the precision of perceived content and produce exact outcomes. This exploration chipped away at two distinct ideas, first is idea of Pixel-level thresholding handling and another is Otsu Method thresholding. With all such images, the algorithm correctly separated the texts even with such images, the number of text and header separation was done successfully and have compared our result with line pixel thresholding technique and Otsu techniques. We have found Otsu method is 30.62% faster than the Line pixel thresholding technique.

REFERENCES