Robust Document Image Binarization Using Degraded Document Images

#1Shruti Barmecha, #2Pooja Bora, #3Dipika Suryawanshi, #4Aparna Thorat

#1shrutibarmecha.94@gmail.com, #2bpooja07@gmail.com, #3dipika3695@gmail.com, #4aparnathorat5@gmail.com

#1234Computer Department,Savitribai Phule Pune University. BVCOWE, Katraj, Pune-411043, Maharashtra, India.

ABSTRACT

In the field of Image Processing Binarization of the image is becoming more efficient due to various newly developed techniques. The OCR is used to recognize the character and store it in the text format. The goal is to binarize the degraded document and recognize the character. In our proposed system we are trying to enhance the efficiency of binarization and OCR by combination of most efficient algorithm for binarization and OCR. We are using RGB to Greyscale, Local thresholding, Manual thresholding, Otsu’s Thresholding for binarization of document image. If any noise remains after binarization the Blurring is used to remove the noise. Along with the binarization, we are providing OCR technique to recognise the content of the document image and save the content in text format. In the OCR we are using Thinning, Segmentation, Scaling, Template Generation and Template Matching. The text format of the content will help in preserving the data. Because of OCR technique human errors and time consumption is reduced. After the process the testing will be conducted on three public datasets that are used in the recent document image binarization contest (DIBCO) 2009 & 2011 and handwritten-DIBCO 2010

Keywords— Binarization, Degraded document Image, Local thresholding, Otsu’s thresholding, OCR, DIBCO

I. INTRODUCTION

Document Image Binarization is a technique which is performed in the preprocessing stage for document analysis and it aims is to extract the foreground text from the document background. A fast and accurate document image binarization technique is important for the ensuing document image processing tasks such as optical character recognition (OCR). Since, many years document image binarization has been studied but, the thresholding of degraded document image is still an unsolved problem because of document background across different document images. Due to these different types of document degradations it tend to induce the document thresholding error and make it a big challenge to most state-of-the-art techniques. A document binarization technique that extends our previous methods. The method is simple, robust and capable of handling different types of degraded document images with minimum parameter tuning. To convert the content into text format it makes use of different thresholding technique for binarization and different OCR technique. In particular, the binarization with OCR technique can reduce the human error and time consumption. Thus, making the project more efficient.

II. Modules.

1. Pre-Processing.
   A. RGB Separation
   B. Grayscale.

2. Binarization.
   A. Thresholding.

3. Post Processing.
   A. Blurring.

4. OCR Pre-processing
   A. Thinning.
   B. Segmentation.
C. Scaling.

5. Template Generation and Matching.

III. Description of Modules

1. Pre-Processing.
   A. RGB Separation.
   B. Greyscale.

Greyscale is used to convert the coloured image into black and white. It consists of shades of grey other than black and white. It is usually required for image processing.

RGB to Greyscale conversion steps are as following.

Algorithm

1. Traverse through the whole input image.
2. Read each pixel color values (24-bit).
3. Find out individual pixel R, G and B 8bit values.
4. Using the formula calculate the greyscale value for given R, G and B values.
5. From 8-bit greyscale value compose 24 bit pixel values.
6. Store the new value in the output image at the same location.
7. Traverse through the whole image.

Now, for computing the greyscale component we use the following steps:

\[ gs = \frac{(r + g + b)}{3}; \]

1. Here the average of R, G, and B is calculated and then saved in the output image.

Formula can also be written as

i. \[ gs = r * 0.33 + g * 0.33 + b * 0.33; \]

2. i.e. 33% of all colors is used to compose final 100% greyscale component.

3. NTSC Formula

ii. \[ gs = r * 0.33 + g * 0.56 + b * 0.11; \]

4. The NTSC formula takes 56% of green where as only 11% of blue since green color has more luminescence or brightness than blue.

2. Binarization.

A. Thresholding: Thresholding is one of the simplest method of image segmentation. Thresholding is used to create binary image i.e. black and white from a greyscale image. It is used where the required features are converted to Black and everything else in white (or vice-versa) i.e. for feature extraction.

Algorithm

1. Traverse through the entire image pixel by pixel.
2. Read each pixel color values (24bit) and then convert into greyscale.
3. Based on current Threshold calculate the binary output pixel value (24 bit).
4. Store the new calculated values at the same location in the output image.

3. Post-Processing.

A. Blurring: In blurring each pixel in the source image gets mixed into the neighbourhood pixel i.e. surrounding pixel.

Algorithm

1. Traverse through entire input image pixel by pixel.
2. Read every pixel color value (24-bit).
3. Split the color value into individual R, G and B 8-bit values.
4. Calculate the RGB average of surrounding pixels and assign this average value to it.
5. Store the new value at same at same location in output image.

\[ Rsum = Rsum + R \]
\[ Gsum = Gsum + G \]
\[ Bsum = Bsum + B \]

\[ R = \frac{Rsum}{(depend\ on\ type\ of\ window\ used)} \]
\[ G = \frac{Gsum}{(depend\ on\ type\ of\ window\ used)} \]
\[ B = \frac{Bsum}{(depend\ on\ type\ of\ window\ used)} \]

4. OCR Pre-processing.

A. Thinning: The basic method for skeletonization is thinning. It this technique, it extracts the skeleton of an object as a result. In every iteration, the edge pixels having at least one adjacent background point are deleted. All those pixels can be removed, only if its removal doesn’t effect the object. It represents the shape of the object in small number of pixels.

It uses a set of four 3 *3 templates to scan the image.

![Templates to identify pixels to be eroded in the Stentiford Method.](image_url)

Figure 1. Shows these four templates.

Figure. Templates to identify pixels to be eroded in the Stentiford Method. The empty white boxes belong to places where the color of the pixel does not need to be checked.

The Stentiford Algorithm can be stated as following:

- Find such a pixel location (i, j ) that matches with the above template T1. Pixels are removed moving from left to right and from bottom to top of the image with this template.

- Mark the pixel for deletion, if the central pixel is not an endpoint, and has connectivity number=1.
Endpoint pixel: The given pixel is endpoint is it is connected to just one other pixel. That is, if a black pixel has only one black neighbour out of the eight possible neighbours.

Connectivity number: It is a measure of how many objects are connected with a particular pixel.

where:

\[ C_n = \sum_{k \in S} N_k - (N_k \cdot N_{k+1} \cdot N_{k+2}) \]

\( N_k \) is the color value of the eight neighbours of the pixel analysed. \( N_0 \) is the central pixel. \( N_1 \) is the color value of the pixel to the right of the central pixel and the remaining are numbered in counter clockwise order around the centre.

\( S = \{1,3,5,7\} \)

Figure illustrates the connectivity number.

1. If all the locations matches with T1 then repeat steps 1 and 2.
2. Repeat steps 1-3 for the rest of the templates: T2, T3, and T4. T2 will match pixels on the left side of the object, moving from bottom to top and from left to right. T3 will select pixels along the bottom of the image and move from right to left and from bottom to top. T4 locates pixels on the right side of the object, moving from top to bottom and right to left.
3. Set to white the pixels marked for deletion.

B) Segmentation and Scaling: After segmentation we get the output as segment of characters i.e. a single character from the word. For the segmentation we will use the Raster scale algorithm. Here the given whole text image is divided into strips known as “scale line”.

Several lines of text may be present in the given input document that needs to be categorized into single character for recognition. For this purpose the following steps are to be applied:

1. The document is to be scanned from the first dark pixel and named as top of row.
2. Now for bottom the next blank line is detected. The area between this is row of characters in image.
3. Now each and every character is recognized from the row. This the done by scanning the row from left to right till the bottom. Then the first dark pixel is the leftmost pixel of character.

4. Now if all pixel are found to be blank then this is right of character.

```
  The document is to be scanned from the first dark pixel
```

![Figure 3. Boundary](image)

- The character segmented is then scaled to the from any pixel size to 15 X 15 pixel.

```
  The document is to be scanned from the first dark pixel
```

![Figure 4. Scaling](image)

5) Template Generation and Matching.

Template Generation:

When the document is put to visual recognition, it may contain handwritten or printed text. With the same or different font or script. Document may contain the information that is not required for example it may contain colors or pictures that are not required or does not give any useful information in the instant sense of character recognition. It may happen that the given character appears in the same document number of times. Here, each character is submitted to the system for recognition and if the same character occurs then it is taken from the system itself. Now, the cropped character of 15 X 15 size is digitalized into array of 15 X 15, where black represents 1 and white 0 shown in figure.

```
In this process, the input image is sampled into a binary window which forms the input to the recognition system. In the above figure, the alphabet A having a single color, either black or white has been digitized into digital cells. This information is to be encoded in a form meaningful to a computer. For this, we assign a value 0 to each white pixel and +1 to each black pixel and create the binary image matrix I which is shown in the figure 5. Digitization of an image into a binary matrix of specified dimensions makes the input image invariant of its actual dimensions. Hence an image is transformed into a binary matrix of fixed pre-
```
determined dimensions irrespective of the size of the image. This stored patterns as they move through the recognition system and establishes uniformity in the dimensions of the input.

**Template matching:**

Pattern based recognition require matching of generated binary format with the existing template for this purpose the binary has been divided into 5 tracks and each track subdivided into 8 sectors. A corresponding track-sector matrix is to be generated, identifying number of pixels in each region.

![Figure 6. Division into tracks and sectors](image)

This can be done using following procedure

1. Firstly we should identify centre of matrix
2. Calculate radius we can say it as, rad by finding pixel with maximum distance from centre using distance formulae.
   \[
   \text{Dist} = \sqrt{((y2 - y1)^2 + (x2 - x1)^2)}
   \]
3. Calculate (rad ÷ 5) to identify size of each imaginary track.
4. Identify imaginary sectors.
5. Generate track-sector matrix by calculating number of 1’s in each intersection of sector and track.

The track-sector matrix generated above is then matched with existing template. The existing template consists of each track-sector intersection value, each track value and each sector value. If all these parameters are found to match with the template values then the resultant is the character identified. The resultant matrix contains unique value for each font and thus makes it easy to identify each font separately.

**IV Architecture of Proposed System:**

![Figure 7. Architecture of proposed system.](image)

The proposed architecture supports multiple clients. For sake of convince we will consider authorized user i.e. server. It will take the input as the degraded document image then pre-processing is performed on the image i.e. the RGB separation and conversion to greyscale. Then the greyscale image is binarized. If the noise remains then the post-processing is performed. Then we will get the output as the image. Here we will preserve the text from the image through OCR by applying different algorithm i.e. are thinning, segmentation, scaling. Then the templates are generated are saved in the system. If the same template occurs the it is matched and recognized and saved in text format.

**Conclusion**

A document image binarization technique along with OCR is a very efficient technique. In this technique, input degraded document image is first converted to greyscale and the thresholding is applied to remove the noise. Finally, once the initial binarization result is derived, the result can be further improved by post processing. This type of document image binarization technique is suitable for different kind of degraded document images. Then the OCR technique is applied to convert the content of document image in text format. In the OCR technique firstly the thinning, segmentation and scaling is done to have a single letter scaled in a particular format. Secondly, in template generation the digitalization of the letter is done and the character database is updated. Then in template matching, the characters are matched and the output is stored in the string. The template generation and template matching algorithm are better approach than using the ANN.

**References**


[2] “REAL TIME OPTICAL CHARACTER RECOGNITION BASED ON FEED FORWARD NETWORKS” Poonam R. Deokar, Dipali A. Badade,
Deepali B. Chavan, Sharadchandra Pawar College of Engineering, University of Pune, India


Miss. Shruti Barmecha, BE, Computer Department, Bharati vidyapeeth’s college of engineering for women, Pune

Miss. Pooja Bora, BE, Computer Department, Bharati vidyapeeth’s college of engineering for women, Pune.

Miss. Dipika Suryawanshi, BE, Computer Department, Bharati vidyapeeth’s college of engineering for women, Pune.

Miss. Aparna Thorat, BE, Computer Department, Bharati vidyapeeth’s college of engineering for women, Pune.