Image Enhancement Processes for Myanmar Printed Text Recognition

Htwe Pa Pa Win, Khin Nwe Ni Tun
University of Computer Studies, Yangon
hppwucsy@gmail.com,knntun@gmail.com

Abstract

Image enhancement processing step is the basic crucial part of the OCR system as the recognition accuracy of OCR systems greatly depends on the quality of the input text image. Most of the OCR systems, especially Myanmar OCR systems are little effort for preprocessing or image enhancement process. And Myanmar glyphs are complex in shape; combine with many circular, dots and line and pixel distribution to form this scripts are not uniform. Therefore, making the input image to be better is important for our native OCR system. In this paper we propose image enhancement processes suit that, through reducing noise, separating text and background, skew detection and rotation, improves the accuracy of the OCR System for Myanmar printed text. These steps can produce a refined image that is ready for the segmentation or character extraction process of the OCR system.

1. Introduction

An OCR system converts a document image into text format for easy editing, storage, transmission, searching, indexing and integrating into other applications [16]. Some practical application potentials of OCR system are: Reading aid for the blind, Automatic text entry into the computer for desktop publication, library cataloging, ledgering, etc., Automatic reading for sorting of postal mail, bank cheques and other documents, Document data compression, Language processing, Multi-media system design, etc [5].

A typical OCR system consists of the following processes:

Preprocessing
Segmentation
Feature Extraction
Classification
Post processing

Current solutions to the problem of optical character recognition (OCR) have advanced to the point where recognition rates of 99% are common for clean, uniformly formatted text. Unfortunately, the performance of most OCR algorithms degrades very rapidly when even small amounts of noise are introduced into the original document or during the scanning process. In many situations, this increased error rate quickly decreases the return on investment to the point where it is not cost-effective to integrate automated recognition technology solutions. To push this critical point lower and deal robustly with noise, OCR systems often perform some type of image enhancement as a preprocessing step [12]. And every process can affect the recognition accuracy and each step depends on the previous one. For example, an image that includes noise or low quality image can cause improper components by the segmentation process. Therefore, preprocessing steps is the basic important part of all OCR system.

The state-of-the-art OCR engines recognize documents printed in Latin and some Oriental scripts with few errors in each page for high quality images. There is no robust OCR for Burmese scripts of Myanmar language. And the previous Myanmar OCR systems are mostly concerned on handwritten character and less emphasizes on image enhancement processes. The writing style of Myanmar scripts is complex and consist circles, dots and lines with horizontally or vertically. The pixel values are not also uniform for some Myanmar fonts. The wrong pixel content can affect the follow-up process to decrease OCR accuracy. Therefore, making the input image to be better is important for our native OCR system.

Preprocessing operations are usually specialized image processing operations that transform the image into another with reduced noise and variation. It takes in a raw image, reduces noise and distortion and removes skewness of the image thereby simplifying the processing of the rest of the stages.

Traditional preprocessing steps involve:

- Gray Scale converting
- Noise removal
- Binarization
- Skew Detection

Most of the researchers are used at least one or more or all of these steps in any order by using different methods depending on the type of the image for their OCR systems.

The rest of the paper is structured as follow: section 2 discuss related works for the image cleaning process, implementation of the detail proposed preprocessing steps are described in section 3 and we conclude this paper in section 4.

2. Related Works

Although many literature state the various implementation strategies for OCR system, there is little suit for image enhancement process [12].

Dance and Bruce [6] provided for identifying, correcting, modifying and reporting imperfections and features in pixel images that prevent or hinder proper OCR and other document imaging processes. Their invention provides that run length compressed images can be analyzed and corrected directly for improved performance.

Elisa H. et al. [7] explored the effects of different image pre-processing methods for document image binarization. They proved that the binarization method is significant in the binarization accuracy, but the pre-processing also plays a significant role. The Total Variation method of pre-processing shows the best performance over a variety of pre-processing methods. And they conclude that no one binarization algorithm is uniformly best over all possible images, and neither is one pre-processing algorithm.

Megan Elmore and Margaret Martonosi proposed an image preprocessing suite that, through text detection, auto-rotation, and noise reduction, improves the accuracy of OCR analysis by using morphological strategy in [13]. However, they stated that more research is needed for noise deduction techniques, integration of other preprocessing steps and better enhancement of letter shapes for increased legibility.

3. Image Preprocessing Implementation

To implement the image correction process, we gain the following image optically including with some noise and variant.

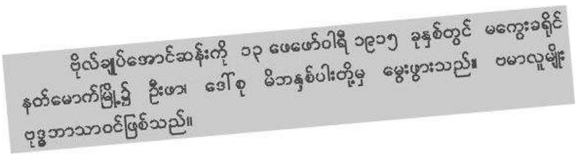
ဗိုလ်ချုပ်အောင်ဆန်းကို ၁၃ ဖေဖော်ဝါရီ ၁၉၁၅ ခုနှစ်တွင် မကွေးခရိုင် နတ်မောက်မြို့၌ ဦးဖာ၊ ဒေါ်စု မိဘနှစ်ပါးတို့မှ မွေးဖွားသည်။ ဗမာလူမျိုး ဗုဒ္ဓဘာသာဝင်ဖြစ်သည်။

Figure 1: Original input image

3.1. Gray Scale Converting

The process of converting color pixel value that is described by a triple (R, G, B) of intensities for red, green, and blue, to a single number of grayscale value. There are many methods for converting to gray scale image. We use luminosity method. The luminosity method is a more sophisticated version of the average method. It also averages the values, but it forms a weighted average to account for human perception. We're more sensitive to green than other colors, so green is weighted most heavily [10]. The effective luminance of a pixel is calculated with the following formula:

Y=0.299RED+0.587GREEN+0.114Blue

For the sake of to make device faster we use the following equation in our program:

```
int Y = ((r * 306) + (g * 601) + (b * 117)) >> 10;
```

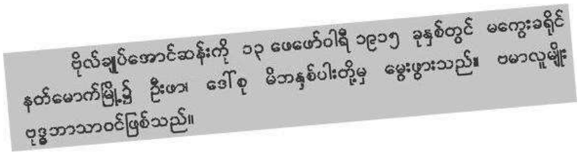
ဗိုလ်ချုပ်အောင်ဆန်းကို ၁၃ ဖေဖော်ဝါရီ ၁၉၁၅ ခုနှစ်တွင် မကွေးခရိုင် နတ်မောက်မြို့၌ ဦးဖာ၊ ဒေါ်စု မိဘနှစ်ပါးတို့မှ မွေးဖွားသည်။ ဗမာလူမျိုး ဗုဒ္ဓဘာသာဝင်ဖြစ်သည်။

Figure 2: Image after converting Grayscale

3.2. Noise Removal

The noise, which is introduced by the optical scanning devices causes disconnected line segments, bumps and gaps in lines, filled loops etc. The distortion which includes local variations, rounding of corners, dilation and erosion is also a problem. Prior to the character recognition, it is necessary to eliminate these imperfections. There are many techniques to reduce the noise which can be categorized in two major groups; filtering and smoothing [9], [15].

Filtering is perhaps the most fundamental operation of image processing and computer vision. In the broadest sense of the term "filtering," the value of the filtered image at a given location is a function of the values of the input image in a small neighborhood of the same location [4].

From the 1980s, many filtering algorithms begun to occur, but it is still very meaningful to remove or furthest restrain the speckle noises when the spatial resolution of the image is not be reduced [20]. Various spatial and frequency domain filters can be designed for this purpose. The basic idea is to convolve a predefined mask with the image to assign a value to a pixel based on its neighboring pixels [14].

In paper [18], they proved that Low Pass Finite Impulse Response Filter is better than other traditional filtering methods in restraining the speckle noises, and the filtering speed is quicker. It has been widely used in digital signal processing and image processing. And for the example of our test image, FIR low pass filter outperform than the median filtering and average filtering methods. Therefore we used FIR low pass filter for our noise removal process.

3.2.1. Implementation principle of the FIR filter

1. First we must establish some parameters for the design such as frequency, Transition width, stop-band attenuation.

2. Calculate the Number of taps or order of the filter:

The number of taps required is normally calculated from the equation

$$N = \frac{3.3}{transition\_width}$$

3. Once the number of taps is known, the filter coefficients can be calculated.

To find the coefficient we use hamming window method and defined the hamming window function

$$w(n) = 0.54 + 0.46\cos\left(\frac{2\pi n}{N}\right)$$, 0<<n<<N

Denote the impulse response function

$$h_D(n) = 2f_c \frac{\sin(n\omega_c)}{n\omega_c}$$, 0<<n<<N

window the h(n) by the selected window function;

$$h(n) = w(n).h_D(n)$$

These $h(n)$ represent the coefficient parameters of the digital filter:

4. Calculate the output of the low pass FIR filter

$$y[n] = \sum_{i=0}^{N} b_i x[n-i].$$

Where,
x[n] is the input signal,
y[n] is the output signal,
bi are the filter coefficients
N is the filter order

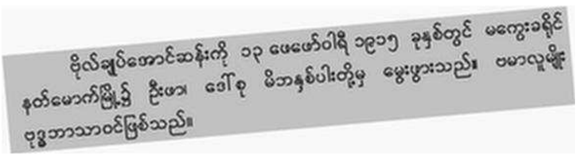
ဗိုလ်ချုပ်အောင်ဆန်းကို ၁၃ ဖေဖော်ဝါရီ ၁၉၁၅ ခုနှစ်တွင် မကွေးခရိုင် နတ်မောက်မြို့၌ ဦးဖာ၊ ဒေါ်စု မိဘနှစ်ပါးတို့မှ မွေးဖွားသည်။ ဗမာလူမျိုး ဗုဒ္ဓဘာသာဝင်ဖြစ်သည်။

Figure 3: Image after removing noise using FIR low pass filter

3.3. Binarization

Binarization, or thresholding, is a conversion from a grey level image to a bi-level image by turning all pixels below some threshold to zero and all pixels about that threshold to one.

If a method computes a threshold for the neighborhood around each pixel or for each designated block in the image, it is called a localmethod. Niblack and Eikvil et al. are the top-ranked local threshold methods in terms of the error rate and rejection rate for character recognition. If a binarization method computes a threshold for an entire image, it is called a global method. Many researchers evaluated four such methods and concluded that Otsu's approach outperforms the other three [3].

The Otsu method also appears to best model the truth behind historical printed documents, based on results over a broad range of historical newspapers. The plugin for most image processing tools also use Otsu thresholding technique. And it can find the threshold automatically based on the input gray data. It is a simple but effective tool to separate objects from background [19]. Otsu threshold is used in many applications from medical imaging to low level computer vision. It has many advantages and assumptions.

Advantages

- Speed: Because Otsu threshold operates on histograms (which are integer or float arrays of length 256), it's quite fast.
- Ease of coding: Approximately 80 lines of very easy stuff.

The limit of this method is that it applicable only when the image is bimodal (the histogram contain two peaks).

3.3.1. Finding the threshold with Otsu Method

Where q1 and q2 represent the estimate of class probabilities defined as:

$$q_1(t) = \sum_{i=1}^{t} P(i) \quad \text{and} \quad q_2(t) = \sum_{i=t+1}^{I} P(i)$$

and sigmas are the individual class variances defined as:

$$\sigma_1^2(t) = \sum_{i=1}^{t} [i-\mu_1(t)]^2 \frac{P(i)}{q_1(t)} \quad \text{and} \quad \sigma_2^2(t) = \sum_{i=t+1}^{I} [i-\mu_2(t)]^2 \frac{P(i)}{q_2(t)}$$

and the class means:

$$\mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{q_1(t)} \quad \text{and} \quad \mu_2(t) = \sum_{i=t+1}^{I} \frac{iP(i)}{q_2(t)}$$

Here, P represents the image histogram. The problem of minimizing within class variance can be expressed as a maximization problem of the between class variance. It can be written as a difference of total variance and within class variance:

$$\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = q_1(t)[1-q_1(t)][\mu_1(t)-\mu_2(t)]^2$$

Finally, this expression can safely be maximized and the solution is t that is maximizing $\sigma_b^2(t)$.

This idea yields an effective algorithm as follow [6].

Algorithm

- Compute histogram and probabilities of each intensity level
- Set up initial ωi(0) and μi(0)
- Step through all possible thresholds $t = 1 \ldots$ maximum intensity
    - Update ωi and μi
    - Compute $\sigma_b^2(t)$
- Desired threshold corresponds to the maximum $\sigma_b^2(t)$

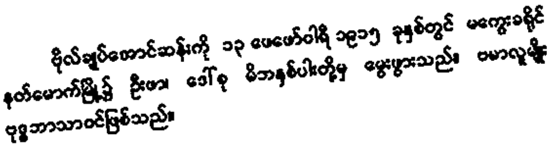
ဗိုလ်ချုပ်အောင်ဆန်းကို ၁၃ ဖေဖော်ဝါရီ ၁၉၁၅ ခုနှစ်တွင် မကွေးခရိုင် နတ်မောက်မြို့၌ ဦးဖာ ဒေါ်စု မိဘနှစ်ပါးတို့မှ မွေးဖွားသည်။ ဗမာလူမျိုး ဗုဒ္ဓဘာသာဝင်ဖြစ်သည်။

Figure 4: Image after using Otsu threading

3.4. Skew Detection and Correction

The scanned document may be slant because of the human error that is misplacement of the paper document on the scanner. This can reduce the OCR performance. Even a smallest skew angle existing in a given document image results in the failure of segmentation of complete characters from words or a text lines, as the distance between the character reduces. Further, most of the OCRs and document retrieval/ display systems are very sensitive to skew in document images. Hence it is important to detect and correct skew because it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages [1],[8].

The process of attempting to estimate the orientation angle, the skew angle, of the text lines is called skew detection. The process of rotating the document with the skew angle, in the opposite direction, is called skew correction. The most popular algorithm for detecting skew is the Hough transformation. Generalized Hough Transform (GHT) is the extension of the standard HT to detect any arbitrary object in an image and can use to recognize printed characters in their different shapes [21][17]. We use the generalized Hough transformation to detect the angle of an image's baselines and rotate the image to the correct angle.

The basic algorithm in pseudo code [11]:

```
1. Create a two-dimensional matrix Hough and
   initialize the values with 0
2. for y=0 to Height-1
3.   for x=0 to Width-1
4.     if Point(x,y) is black then
5.       for alpha=-20 to 20 step 0.2
6.         d= Trunc(y*cos(alpha)-x*sin(alpha))
7.         Hough (Trunc(alpha*5),d)+=1
8.       next alpha
9.     end if
10.  next x
11. next y
12. Find the top 20 (alpha,d) pairs that have the
    highest count in the Hough matrix
13. Calculate the skew angle as an average of the
    alphas
14. Rotate the image by – skew angle
```

The advantages and disadvantages of generalized Hough transform is described in [2] as follow:

Advantages

-The generalized Hough transform is essentially a method for object recognition.

-It is robust to partial or slightly deformed shapes (i.e., robust to recognition under occlusion).

-It is robust to the presence of additional structures in the image (i.e., other lines, curves, etc.).

-It is tolerant to noise.

-It can find multiple occurrences of a shape during the same processing pass.

Disadvantages

-It requires a lot of storage and extensive computation (but it is inherently parallelizable!).

-Faster variations have been proposed in the literature:

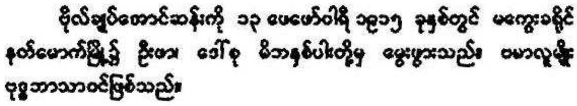
ဗိုလ်ချုပ်အောင်ဆန်းကို ၁၃ ဖေဖော်ဝါရီ ၁၉၁၅ ခုနှစ်တွင် မကွေးခရိုင် နတ်မောက်မြို့၌ ဦးဖာ ဒေါ်စု မိဘနှစ်ပါးတို့မှ မွေးဖွားသည်။ ဗမာလူမျိုး ဗုဒ္ဓဘာသာဝင်ဖြစ်သည်။

Figure 5: Image after correction of skew using Generalized Hough Transform

4. Conclusion

We described the way for image enhancement processes as a first step for our Myanmar printed OCR system. After this process, we can initiate the next process of Segmentation where we isolate the character form the image outputted from the preprocessing step. The good way of image enrichment process may prevent the occurrence of the wrong segmentation and then may increase the overall accuracy of the OCR system.

5. References

[1] *A. M. Al-Shatnawi and K. Omar* , "Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity", *Journal of Computer Science 5 (5):,* Science Publications,2009, ISSN 1549-3636, pp. 363-368

[2] Ballard and Brown, Sonka et al.,Generalized Hough Transform (GHT)

[3] C. H. Chou, W. H. Lin, FuChang, "A binarization method with learning-built rules for document images produced by cameras", *Pattern Recognition, Vol 43,* Academia Sinica, Taipei, Taiwan, 2010, pp.1518-1530

[4] C. Tomasi _ R. Manduchi, “Bilateral Filtering for Gray and Color Images”, *Proceedings of the IEEE International Conference on Computer Vision*, Bombay, India, 1998.

[5] D. Achaya U, N. V. Subba Reddy and Krishnamoorthi, “Hierarchical Recognition System for Machine Printed Kannada Characters”, *IJCSNS International Journal of Computer Science and Network Security*, VOL.8 No.11, November 2008.

[6] D. R. Allan and B. B. Fast, “ OCR Image Pre-Processor”, United State Patient, No. 5778103, July 7, 1998.

[7] E. H. B. Smitha, L. L. Sulemb and J. Darbon, “Effect of pre-processing on binarization”, Boise State University, Boise, Idaho, USA; Telecom ParisTech, Paris, France, UCLA, Los Angeles, California, USA, 2010.

[8] G. H. Kumar, Nandini N. and Srikanta M. K., “Estimation of Skew Angle in Binary Document Images Using Hough Transform”, *World Academy of Science, Engineering and Technology* 42 2008.

[9] Gonzalez, R., Woods, R., *Digital Image Processing, 2nd ed*., Prentice Hall, 2002.

[10] John, “Three algorithms for converting color to grayscale”, August 24, 2009.

[11] Mackenb, “how to deskew an image”, 25 Apr 2006.

[12] M. Agrawal, D. Doermann, J. Kumar, W. A. Almageed, “Document Image Enhancement”, Monday 30 August, 2010.

[13] M. Elmore and M. Martonosi, “A Morphological Image Preprocessing Suite for OCR on Natural Scene Images”, Georgia Institute of Technology, Princeton University, 2008.

[14] M. S. M. El-Mahallawy, “A Large Scale Hmm-Based Omni Font-Written Ocr System For Cursive Scripts”, Cairo University Giza, Egypt, April 2008.

[15] Pratt, W.K., *Digital Image Processing, 2nd edition*, John Wiley & Sons Inc., 1991.

[16] S. K. Shah and A. Sharma, Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching", *Gujarat.IE(I) Journal−ET*, 2006.

[17] S. Touj, N. B. Amara, and H. Amiri, “Generalized Hough Transform for Arabic Printed Optical Character Recognition”, *The International Arab Journal of Information Technology, Vol. 2, No. 4,* October 2005.

[18] W. Zhang & J. Yang, C. Fang, D. Fan, J. Wu, “Noise Removal of Spaceborne SAR Image Based on the FIR Digital Filter”, *Computer and Information Science, Vol. 3,* No. 1, February 2010.

[19] Xu Liang, “Image Binarization using Otsu Method” , NLPR-PAL Group CASIA, April 20th, 2009.

[20] Zhang, Wei, Sun, Ronghua & Zhang, Xiaocan, “Noise Removal of SAR Image Based on Improved Wavelet Soft-thresholding Method”, *Remote Sensing Information.* No.4. P.4-6. 2006.

[21] www.wikipedia.com