Handwritten Myanmar Character Recognition System using the Otsu's Binarization Algorithm

Aye Aye Nyein[1], and Hlaing Htake Khaung Tin[2]
[1]Faculty of Computer Science, University of Computer Studies, Hinthada, Myanmar
ayeayenyein2008@gmail.com
[2]Faculty of Infomation Science, University of Computer Studies, Hinthada, Myanmar
hlainghtakekhaungtin@gmail.com

Abstract. Handwritten is a famous techniques involved in many recognition system in image processing and pattern recognition. In this paper, first step is pre-processing includes noise reduction, colour conversion and edge detection method to segment Handwritten Myanmar Characters from the image. Also, Open Source Tesseract OCR engine is used to recognize the segmented Handwritten Myanmar Characters. The recognition accuracy will highly depend on the good segmentation. Some of the handwritten recognition applications are criminal documentation, security method, image and Movie managing and medical science research on Myanmar character recognition is still in its early days with limited literature available till date.

Keywords. Handwritten, myanar character, recognition system, feature extraction, edge detection.

1 Introduction

Handwriting recognition is also recognized as handwriting recognition. The computer is used to access and restrict sensitive handwriting such as paper forms and other tools. The source of the written text can be seen on a piece of paper. Or it can be recognized by writing. Otherwise, the function of the pen tip is set to "on line". For sample, a pen-based computer display will appear. It's usually easier. The handwriting recognition system converts guess work into practical letters and finds the most probable words. [1]

Off-line handwriting identification includes routine translation of manuscript in the image and changes in the text and code that works on your computer. The information obtained in this form is representative of the handwriting. Off-line handwriting identification is challenging, as there are different types of handwriting. Today, the OCR engine focuses on ICR for handwriting and typography.

On-line handwriting recognition is written in a special digitizer or PDA, which automatically changes to writing. It has a sensor that collects pen-top movements and pen-down moving. This information is known as digital ink and can be regarded

ICIDSSD 2020, February 27-28, New Delhi, India
Copyright © 2021 EAI
DOI 10.4108/eai.27-2-2020.2303219

as a hand-held digital statement. The received signal is changed to computer alphabet and text alphanumeric code.

Recognizing well-typed words in contemporary print is a major problem. Different types of handwriting in handwriting. Many of the papers used daily are written documents such as postal addresses, Bank checks; Prescriptions; Many historical documents, And important components of information collected through forms. In many cases, these documents may be indexed, indexed, or indexed. It is interesting to have these documents in digital form, so as to provide new ways of consulting and working.[4]

Generally, handwriting recognition is divided into two categories: classification and online handwriting recognition methods. On off-line recognition, scans are usually taken well by scanners and the final image is available as a photo. However, on-line systems provide two-dimensional coordinates of coordinates as a function of time, and can also be performed by the author. Handwriting recognition is one of the most interesting and challenging research areas in the field of imaging and pattern recognition in recent years. Comparison of neural networks in offline letter recognition systems. High levels of accreditation are used successfully.[5]

Handwriting recognition means the paper, paper, and computer files of a computer. Photos get handwritten information from sources such as the touch screen and other products. Optical reading from a piece of optical character paper written. [6]

2 Literature Review

Handwritten Myanmar Character Recognition based on Competitive Neural Trees (CNet) proposed in [14]. They worked the steps of binarization, thinning and resizing in pre-processing. Region-based methods are used to extract features. Eighteen features are extracted. In the recognition stage, Competitive Neural Trees (CNeT) is used. CNeT consists of five steps: CNeT learning phase, the life cycle of CNet, the training procedure of CNet, recall procedures and global search method.

Passport automatic entry system for security system. Skewing, resizing, filtering, grayscale converting, normalization steps are done in pre-processing steps. For segment, use the region-based segment. The shape and skeleton has been expanded for the production of techniques.[7]

A model to recognize the payee's name and legal amount, the courtesy amount on a bank cheque. For pre-processing, the noise removing, cropping, skewing, detection, thinning and normalization methods are applied. In the feature extraction step is used MWR algorithm. They classified Myanmar character into many groups depend on the nature of its writing style.[1]

3 Pre-processing steps of handwritten character recognition system

It is necessary to eliminate relevant information for input figures that may affect the decision of preprocessing. Often includes preprocessing binarization, normalization and selection, smoothing and denoising. It is the second step to remove the feature. The preprocessing algorithms take two or more field-scale vulnerabilities available to the field with higher spectral data. The online handwriting identification process can be cracked down into a few simple steps.

- pre-processing
- feature extraction
- categorization or classification

3.1 Pre-processing

This process includes noise reduction, colour converts to grey scale image, and edge detection of facial images.

3.2 Noise Reduction

On the camera or on a magnifying glass Scanners, such as imperfections in the scanner light, indicate the noise in the scanned face.

3.3 Noise Reduction using Media Filter

The average filter is a nonlinear digital filtering method. Frequent noise can be removed. It is a fundamental building block in many situations of visual change. However, its use has long been an algorithmic complication for the kernels radius. A more effective median filtering algorithm is needed, with a focus on the bigger picture. It shows higher angle or higher accuracy data and proximity to a circular kernel. The average filter is better at removing the noise. Reducing such noise remains a standard pre-correction phase to progress the performance of later actions (eg, scanning on an image).

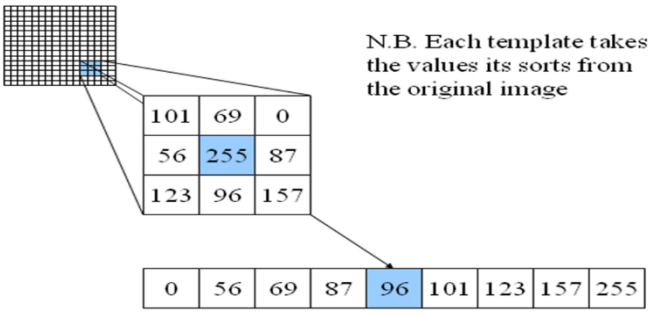

Fig.1. Noise reduction using with median filter

3.4 Convert Grayscale Image

RGB image converts to gray. The gray scale images are different from the black-and-white images of bacteria that match the image of a computer. These are only two images: black and white (also known as either level or decimal).

Between the gray scale images there are many shades of gray. They are also called monochromatic and refer to the presence of a single chrome. They are often the result of measuring the intensity of light in an individual pixel in an electromagnetic spectrum (eg, infrared, visible light, etc.).

3.5 Edge Detection

Edge detection refers to the error-finding and search operation in the image. Degradation is a sudden change in the intensity of a pixel that indicates the boundary of the object in the field of view. Normal methods add include direct injection of the image to the 2-D filter. It is designed to highlight large gradients from the image and return zero values to the uniform region.

Edge detection refers to the image detection and recovery process. Degradation is a sudden change in the intensity of a pixel that indicates the boundary of the object in the field of view. Normal filtering methods include directly injecting an appearance into a 2-D filter. It is designed to be searching to the great gradients of the appearance and returns the value to zero.

Horizontal Operators can use operators to find vertical or horizontal edges. Since both noise and edge are high-frequency, Edge detection is difficult for noisy images. Efforts to reduce noise have been blurred. Get a bad break. Operators used on noisy images are generally broad. As a result, they have sufficient information to reduce local noise levels. This results in a specific version change of the edge found. No step adjustment for all edge intensities.

Effects such as light refraction or low concentration sometimes cause objects in the boundaries defined by the changes. In such cases, the operator needs to make a choice. Therefore, false discovery and missing symptoms, Losing the true edges.

4 Proposed System of Handwritten Character Recognition System

Handwriting recognition is a popular and well-known research area. An overview of the proposed system design is shown in Figure 2 below.

The first step of this system is accepting the input image that contains any Handwritten Myanmar Characters. Then, the system converts the input image to grey scale image. And then, the system converts the resulted grey scale image to binary image using Otsu's binarization algorithm. Then, the system reduces noises by using median filter method.

Average filtering is a nonlinear technique cast-off to eliminate noise from imagination. It is widely used. This is because it is very effective in removing root canals and noise. It especially removes salt and pepper.

The average filter mechanism by progressing the image pixel by pixel by substituting each value with the average value of neighboring pixels. Neighborhood patterns are called "windows". It slows down the whole picture. Pick it up. The average is calculated by first calculating the pixel standards in the space by first order, then replacing the average pixel value and replacing the calculated pixel.

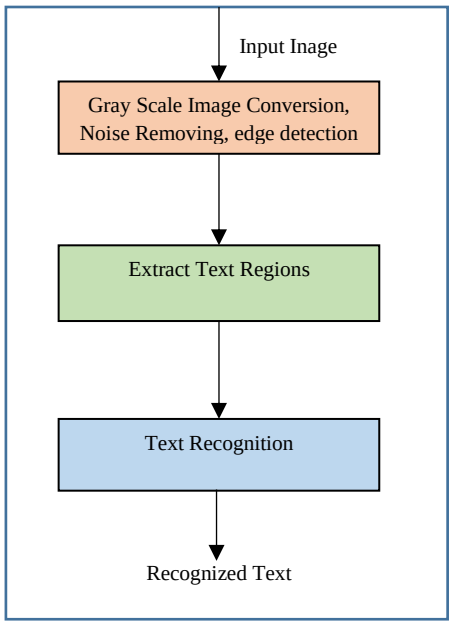

Fig.2. Overview of the proposed system design

The categorized image $S = \{S(i,j)\}$ from SMF can be described by the following equation:

$$S(i,j) = Median\ (k,l) \in\ W_{m,n}\{D(i+k, j+1\}(1)$$

Where $W_{m,n}$ is a sliding window of size m *n pixels centered at coordinates (i, j). The median rate is evaluated by operating equation (1) with ns = m*n [8].

The system then uses the Canny Edge detection algorithm to find edges in binary images. In general, the purpose of the edge detection is to decrease the sum of

information in an image and to maintain the physical assets used for more image processing.

The Canny Edge algorithm is a multi-step algorithm with the following steps:

- Smoothing
- Gradient discovering
- Non-maximum Suppression
- Double Thresholding
- Edge Tracking

Then, the resulted edge image is dilated using cross kernel. The dilation operation thickens the image. The extent of how much it should be thicken is based on the structuring element [2].

	1	
1	1	1
	1	

Fig.3. Cross Kernel

Then, the system finds text regions using connected component labelling method. For connected component component labelling, classical Rosenfeld and Pfaltz algorithm is used. The algorithm is two pass algorithm using four neighbourhood.

	1	
2		3
	4	

Fig.4. Four neighbourhoods

In the first phase, the procedure drives through each pixel. It inspections the pixel above (North) and to the left (West). And using these pixel's labels (which have already been assigned), it allocates a label to the current pixel. In the second phase, assign the temporary label with its equivalent label.

Algorithm is as follows [8]:

(step -1) From left to right Scan the image from top to bottom.

(step -2) If the pixel is a foreground pixel, then

(2.1) If it has a logo in its upper and left neighbor, copy the label.

(2.2) If both have the same logo.

(2.3) If both labels have different labels, Copy the marker above.

(2.4) Otherwise, set this pixel to a new label and add it to the Equation Table.

(step -3) If there are no pixels to consider, go to step 2.

(step -4) Find the lowest common denominator for each of the equations in the Equation Chart.

(step -5) Scan image Replace each brand with the lowest one.

After connected component labelling, each component can be assumed as a text region and extracted. Finally, each text region is recognized with the open source tesseract OCR engine and the system displays the recognized text.

The determination of this phase is to highpoint the significant material for the accredited model. This facts includes the weight of the pen. Information may include information such as changes in writing or text. The last big step is segregation. At this stage, the characters or words used by mapping the features of different models to different classes. [2]

5 Results and Discussion

The more content you have, the better. A better system output will result. To improve the accuracy of the system's results, the system must use appropriate content extraction and classification methods.

Developing OCR for handwritten characters is quite challenging and prone to errors due to structural complexity. Learn how to find the best available method. Although methods for obtaining better results are slower in nature, fasting methods are generally more impractical, and it is found that the OCR method provides more accurate results than other methods. In addition, accuracy rate is dependent upon pre-processing of the system. The example of input image is shown in the resulting figure 5.

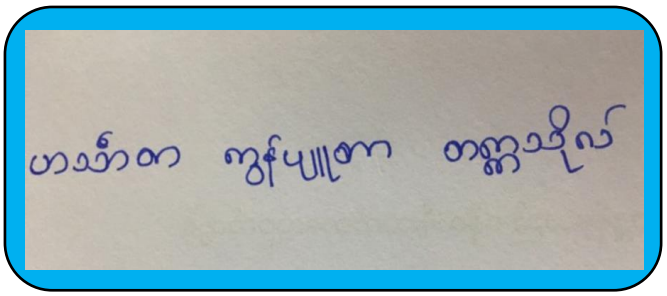

Fig.5. Example of Input Image

And the detected text regions are presented in the resulting figure 6.

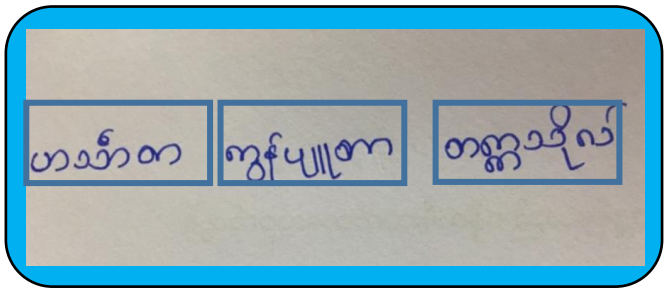

Fig.6. Detected Text Regions

6 Conclusions

This paper is presented to utilize canny edge detection and connected component labeling to provide good segmentation results of Myanmar handwritten characters. Also, the system uses neural network based open source Tesseract OCR engine to recognize the segmented Myanmar handwritten characters. This paper use recognized Myanmar words count and unrecognized Myanmar words count to determine the accuracy of the system. Testing input pictures are used in 300 dpi, 400 dpi, and 500 dpi images of the handwritten Myanmar characters. By the experimentation results, 300 dpi images is appropriate for good recognition rate of handwritten Myanmar characters. By the experimentation results of the system shown in figure 6, there are 2 unrecognized Myanmar words and 7 recognized Myanmar words count of total 9 Myanmar words. So, nearly 78 % of the Myanmar words can be recognized accurately. When the system is experimented with other Handwritten Myanmar Characters, below 75% and around 85% accuracy rate are available. The advantages of using OCR systems are reducing the data entry time and requiring the storage space by documents.

References

[1] https://en.wikipedia.org/wiki/Handwriting_recognition

[2] Huang, B. Zhang, and Kechadi, Preprocessing Techniques for Online Handwriting Recognition, Intelligent Text Categorization and Clustering, Springer Berlin Heidelberg, 2009, Vol. 164, "Studies in Computational Intelligence" pp. 25–45.

[3] Holzinger. A.; Stocker. C.; Peischl. B. and Simonic. K.M.; On Using Entropy for Enhancing Handwriting Pre-processing, Entropy 2012, pp. 2324-2350.

[4] Ms. Rohini R. Varade, "A Survey on Various Median Filtering Techniques for Removal of Impulse Noise from Digital Images.", (IJARCET) Volume 2, Issue 2, February 2013.

[5] J.-G. Lu, "Title of paper with only the first word capitalized," J. Name Stand.

[6] Pratibha A. Desai, "Design and Simulation of Handwritten Text Recognition System", International Journal of Current Engineering and Technology, 2013.

[7] M. M. Thinn, "Hand Written Recognition System for Automatic Data Entry of Passport", Mandalay Technological University, Myanmar.

[8] Luigi .D. S, Andrea. B, "A Simple and Efficient Connected Components Labeling Algorithm", University of Bologna, Italy.