2010 Third International Conference on Knowledge Discovery and Data Mining

A Review of Evaluation of Optimal Binarization Technique for Character Segmentation in Historical Manuscripts

Chun Che Fung and Rapeeporn Chamchong
School of Information Technology
Murdoch University
Perth, Western Australia
l.fung@murdoch.edu.au, r.chamchong@ieee.org

***Abstract*— A number of binarization techniques have been proposed in the past for automatic document processing. Although some studies have aimed to evaluate the performance of binarization algorithms, there is no automatic system that is capable of selecting the most appropriate method of binarization. While preprocessing techniques can be applied, binarization is essential to extract the objects in the first place before the characters can be separated for recognition. Although there are several commonly used binarization approaches, there is no single algorithm that is suitable for all images. Hence, there is a need to determine the optimal binarization algorithm for each image. The objective of this paper is to present a survey of the existing methods of binarization and evaluation measurement which have been developed recently. This will lead to the proposal and development of an approach for automatic selection of binarization techniques in handling historical document images.**

Keywords- binarization, image Segmentation, evaluation measurement, quantitative measurement

I. INTRODUCTION

In Thai history, In Thailand, there exists a huge collection of ancient manuscripts that have invaluable knowledge about the history, culture, and local wisdom of Thai civilization. Many of these documents are recorded on media such as palm leaves or papers in very primitive forms. These documents are deteriorating due to age and lack of preservation facilities at the place of collection

In present, computer technology can process a large amount of images of these documents in multimedia formats for future analysis and storage. Although current systems can store all these images, there is no specific system that is capable to retrieve relevant information efficiently and to extract knowledge from them. It is therefore a key objective of this study to develop an efficient image processing system that could be used to retrieve knowledge and information from these historical manuscripts. However, it is recognized that it is not an easy task as there are many styles of traditional Thai handwriting, noise on the images, and fragmentation or cracks due to fragility of the aged leaves.

It is common that images of the collected historical documents are of poor quality due to insufficient attention paid to the condition of the storage and the quality of the written material. As a result, the foreground and background in the scanned images are difficult to be separated. In this research, palm leaf images are domain data that have varying contrast and illumination, smudges, smear, stains, and ghosting noise due to seeping ink from the other side of the manuscripts. Prior to the stage of knowledge extraction, characters or text on the images have to be recognized. There are three steps which need to be completed prior to the task of character recognition. First, a palm leaf is scanned into a RGB image and then it is converted to a gray-scale image. Next, image enhancement is used to enhance the quality of the image. After this stage, binarization is applied and then text and character separation are carried out before character recognition.

Binarization is an essential part of the preprocessing step in image processing, converting gray-scale image to binary image, which is then used for further processing such as document image analysis and optical character recognition (OCR). Consequently, both image enhancement and binarization of historical document are crucial to remove unrelated information, noise and background on the documents. If these steps are inefficient, the original characters from the image may be lost or more noise may be added. Furthermore, these techniques are essential to improve the readability of the documents and the overall performance of the process.

Several binarization algorithms have been proposed in [1-10]. However, it is difficult to select the most appropriate algorithm. The comparison of image qualities from those algorithms is not an easy task as there is no objective evaluation process to compare the results. In contrast, some researchers have proposed a quantitative image measurement of binarization. This performance evaluation of binarization algorithms is recognized to significantly depending on the image content and on the methodologies of binarization. The common approach is to design a set of criteria and scores of criteria. The criteria may be computed by machine or may be decided by visual human.

The purpose of this article is to study the previous research work in evaluation of optimal binarization techniques on historical documents. Section 2 describes the binarization techniques and section 3 explains about measurement of image quality. Section 4 describes the proposed framework for automatic selection of an optimal binarization algorithm. Finally, a discussion on future research is given in the last section.

II. BINARIZATION TECHNIQUES

Binarization is the task of converting a gray-scale image to a binary image by using threshold selection techniques to categorize the pixels of an image into either one of the two classes. Most of studies [1-4, 10, 11] separated the binarization techniques into two main

978-0-7695-3923-2/10 $26.00 © 2010 IEEE
DOI 10.1109/WKDD.2010.110

methods that are global thresholding and local adaptive thresholding techniques

1) Global Thresholding Techniques attempt to find a suitable single threshold value (Thr) from the overall image. The pixels are separated into two classes: foreground and background. This can be expressed as follows [1]

$$I_b(x,y) = \begin{cases} black & if \quad I_f(x,y) \le Thr \\ white & if \quad I_f(x,y) > Thr \end{cases} \quad (1)$$

where $I_f(x,y)$ is the pixel of the input image from the noise reduction and $I_b(x,y)$ is the pixel of the binarized image.

Otsu's algorithm [7] is the most popular global thresholding technique. Moreover, there are many popular thresholding techniques such as Kapur and et al [12], and Kittler and Illingworth [13].

2) Local Thresholding Techniques [10] calculate the threshold values which are determined locally based on pixel by pixel, or region by region. A threshold value (Thr(x,y)) can be derived for each pixel in the image, and the image can be separated into foreground and background as given in expression (2) [1].

$$I_b(x,y) = \begin{cases} black & if \quad I_f(x,y) \le Thr(x,y) \\ white & if \quad I_f(x,y) > Thr(x,y) \end{cases} \quad (2)$$

The conventional local adaptive thresholding techniques are algorithms by Niblack [6] and Sauvola [8].

Sezgin and Sankur [5] surveyed many thresholding techniques. The summary was divided into six categories as follows

1) Histogram shape-based methods: examples of these techniques are the "convex hull thresholding" proposed by Rosenfeld [14].

2) Clustering-based methods: some researchers used mean-square clustering which was proposed by Otsu [7].

3) Entropy-based methods: an illustration of this technique can be found in Kapur, Sahoo and Wong [12].

4) Object attribute-based methods: this technique can be found in Tsai [15].

5) Spatial methods: examples of this technique are shown in Pal and Pal [16].

6) Local adaptive methods: as shown by Niblack's [6] and Sauvola's algorithms [8].

Many researchers have applied different thresholding techniques with document images with both printed and handwritten text. Some of those algorithms are more efficient in specific documents. On the other hand, in the authors' pervious paper [1], it was demonstrated that when the thresholding techniques was applied to evaluate ancient Thai manuscripts on palm leaves, no single method could be claimed to give an optimal result for all images. In addition, most decisions on how to choose these algorithms were subjectively decided by human. There is no objective way to decide whether the optimal result has been achieved. Leedham and et al. [4] compared five thresholding algorithms by evaluating the precision and recall value of word in the foreground. J. He and et al. [3] compared six binarization algorithms by using word recognition. Sezgin and Sankur [5] surveyed 40 binarization algorithms and categorized them based on the exploitation of their information content. They measured and ranked by using performance criteria. In the next section, measurements on how to determine the goodness of the result image are described.

III. Measurements of Image Quality

Zhang [17] studied different methods of image measurement for segmentation techniques. These methods can be separated into three groups; the analytical, the empirical goodness and the empirical discrepancy groups.

1) The analytical methods treat the algorithms for segmentation directly by considering the principles, requirements, utilities, complexity, etc., of the algorithms. Although properties of segmentation algorithms can be easily obtained by analysis, other properties cannot be analyzed because no formal model exists.

2) The empirical goodness methods evaluate the performace of algorithms by judging the quality of segmented image with certain quality measures generated according to human intuition. Different types of measures, which are intra-region uniformity, inter-region contrast and region shape, have been proposed to assess the goodness algorithm.

3) The empirical discrepancy methods compare the difference between an segmented image and a ground truth image to evaluate the performance of segmentation algorithms. There are five groups of this methods that are based on the number of mis-segmented pixels, position of mis-segmented pixels, the number of objects in the image, feature values of segmented objects and miscellaneous quantities.

The experiments have shown that discrepancy methods are more effective than the goodness methods. However, these methods have to compare with ground truth image so these methods are more complex than the other methods. One possible means of generate ground truth image is to use synthetic images.

Sezgin and Sankur [5] described different performance criteria for binarization algorithms. They used five performance criteria as shown below.

1) Misclassification error (ME)

$$ME = 1 - \frac{|B_O - B_T| + |F_O - F_T|}{|B_O| + |F_O|} \quad (3)$$

where ME varies from 0 to 1 for a perfectly classified image to a totally wrongly binarized image respectively, B_O and F_O are background and foreground of ground-truth image respectively, B_T and F_T are background and foreground of area pixels in the test image respectively, and |.| is the cardinality of the set

2) Edge mismatch (EMM) is defined as [9]

$$EMM = 1 - \frac{CE}{CE + \omega\left[\sum_{k\in\{EO\}} \delta(k) + \alpha\left[\sum_{l\in\{ET\}} F(l)\right]\right]} \quad (4)$$

where *CE* is the number of common edge pixels found between ground-truth image and the binarized image, *EO* is the set of all excess original edge pixels, *ET* is the set of all excess thresholded edge pixels, ω is the penalty associated with an excess original edge pixel, α is the ratio of the penalties associated with an excess thresholded edge pixel to an excess original edge pixel, and $\delta(k)$ is a distance function shown as

$$\delta(k) = \begin{cases} |d_k| & if\ |d_k| < Maxdist \\ D_{Max} & Otherwise \end{cases} \quad (5)$$

where d_k is the Euclidean distance of the k^{th} excess edge pixel to a complementary edge pixel within a search area determined by $Maxdist = 0.025N$, where $N = \sqrt{N_{hor}.N_{vert}}$, $D_{Max} = 0.1N$, $\omega = 10/N$ and $\alpha = 2$.

3) Region non-uniformity (NU) is defined as [5, 17, 18]

$$NU = \frac{P_f \sigma_f^2}{\sigma^2} \quad (6)$$

where P_f is the foreground class probability, σ_f^2 is the foreground variance and σ^2 is variance of whole image. A well-segmented image will have a non-uniformity measure close to 0.

4) Relative foreground area error (RAE) measure for the area feature F = A as follows [17]

$$RAE = \begin{cases} \dfrac{A_O - A_T}{A_O} & if\ A_T < A_O \\ \dfrac{A_T - A_O}{A_T} & if\ A_T \le A_O \end{cases} \quad (7)$$

where A_O is the area of reference image and A_T. is the area of binarized image. *RAE* is 0 if it is a perfect match of the segmented regions.

5) Shape distortion penalty via Hausdorff distance is expressed as:

$$MHD(F_O, F_T) = \frac{1}{|F_O|} \sum_{f_O \in F_O} d(f_O, F_T) \quad (8)$$

MHDs are calculated for each 19x19 pixel character box and then the *MHDs* are averaged over all characters in a document. The normalized of *MHD* value to the highest *MHD* value over the test image set *NMHD*.

The score of five performance measure for the i[th] image is shown as:

$$S(i) = [ME(i) + EMM(i) + NU(i) + RAE(i) + NMHD(i)]/5 \quad (9)$$

This technique was implemented to measure the quality of 40 thresholding algorithms over two different context of images. Although they found that the clustering-based method of Kittler and Illingworth [13] is the best quality of thresholding techniques in both types of images, they investigated that there is no single algorithm which could be successful for all image types, even in a single domain.

Pavlos and et al [19] surveyed the evaluation of binarization algorithms on historical documents, which was proposed by using statistical measures of image quality description. The evaluation measurement is combined the pixel error rate (*PERR*), the mean square error (*MSE*), the signal to noise ratio (*SNR*), and the peak signal to noise ratio (*PSNR*). The measurement can be described as

$$PERR = \frac{pixelerror}{MxN} \quad (10)$$

$$MSE = \frac{\sum_i \sum_j e(i,j)^2}{MxN} \quad (11)$$

where the local pixel error is $e(i,j) = x(i,j) - y(i,j)$, black and white value are 0 and 255 for gray-scale images respectively, *x(i, j)* and *y(i ,j*) are a pixel of original image and output image respectively, and *MxN* is size of image. Consequently, *PERR* definition will be

$$PERR = \frac{MSE}{255^2} \Leftrightarrow MSE = PERR \bullet 255^2 \quad (12)$$

$$SNR(DB) = 10 \log_{10} \frac{\sum_i \sum_j x(i,j)}{PERR \bullet 255^2} \quad (13)$$

$$PSNR(DB) = 10 \log_{10} \frac{M \cdot N}{PERR \bullet 255^2} \quad (14)$$

They applied the proposed technique to 30 binarization and compared the result image with the original pdf document image by counting changed pixels (white-to-black or vice versa). Their data set was synthesized from a clean document image (doc), which was considered as the ground truth image, and noise was add to original image. They found that even though the local binarization techniques presented a better quality of result, the global technique based on histogram or classification techniques gave as good results as the local technique.

Badekas and Papamarkos [20] proposed the technique to combine the best binarization results from the independence binarization techniques (IBT) such as Otsu, Niblack, Sauvola, and so on by using their best parameter set (PS) and the Kohonen self-organizing map (KSOM) neural network in the final stage. The paper explained that it is not known initially the best result and this is a main problem of the validity of comparison. They used ground truth image to estimate the best result, called as estimated ground truth (EGT), and compared with IBT results using ROC analysis or Chi-square test. The best PS of the best result from IBT is fed to KSOM. After this, the final binary image is produced by combining the binary information from independent binarization technique.

IV. A Framework of an Automatic Selection of Optimal Binarization Algorithm

Recently, there are a few algorithms that target specifically on historical document images. Pavlos and et al. [19] evaluated the binarization techniques on historical documents by adding noise to synthesized images. As a result of this, the ground truth image can be generated easily. However, historical documents in the real world are definitely different, and samples of such images with noisy are showed

in Figure 1. These images, palm leaf manuscripts, were used in [1] and it was found that there is no single technique which is suitable for all images and there are some example results of binarization techniques in Figure 2.

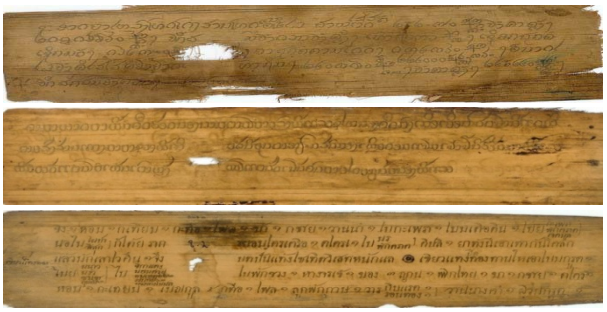
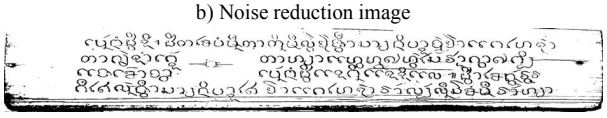


Figure 1. Samples of palm leaf images

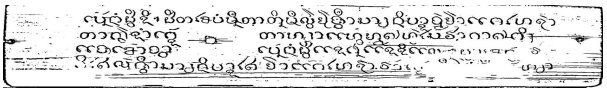
a) RGB image

b) Noise reduction image



c) Binary image by Otsu's algorithm

d) Binary image by Niblack's algorithm



e) Binary image by Sauvola's algorithm

Figure 2. Samples of palm leaf images

This paper proposes a new method to select the optimal binarization algorithm by the following steps:

1. Select appropriate binarization algorithms.
2. Cluster training and testing image sets by using Yitzhaky and Peli [21] for edge detection evaluation to estimate the best binarization technique of each image.
3. Extract features from the noise reduction image by using values from the binarization methods to determine the binarization technique such as histogram of image, total mean and variance of histogram, mean and variance between group of histograms, co-occurrence probability, correlation, means and variances of local areas, entropy values and so on.
4. Classify the optimal algorithm for the image by using machine learning technique.

V. CONCLUSION AND DISCUSSION

From survey, many researchers have evaluated and compared several algorithms by using different measurements. In addition, most of the measurements have to compare the result image with a ground truth image. It is recognized that there is no single binarization technique that is suitable for all images and there is no automatic selection of the optimal binarization technique. This paper proposes a framework to be applied for the processing of ancient manuscripts written on media such as palm leaves. This research will implement and evaluate the proposed optimal binarization technique using machine learning algorithms and features extracted from the binarization process and the noise reduced images. Subsequent development on this work will be reported in the future.

ACKNOWLEDGMENT

The authors wish to thank the Preservation of Palm Leaf Manuscripts Project at Mahasarakham University, Thailand for their support and providing ancient manuscript images.

REFERENCES

[1] R. Chamchong and C. C. Fung, "Comparing background elimination approaches for processing of ancient Thai manuscripts on palm leaves," in *2009 Int. Conf. Machine Learning and Cybernetics,* China, 12-15 July, 2009.

[2] Y. Chen and G. Leedham, "Decompose algorithm for thresholding degraded historical document images," in *IEE Proceeding Visual Image Signal Processing,* December, 2005.

[3] J. He, Q. D. M. Do, A. C. Downton, and J. H. Kim, "A comparison of binarization methods for historical archive documents," in *Proc. 8th Int. Conf. Document Analysis and Recognition,* 2005, pp. 538-542 Vol. 1.

[4] G. Leedham, Y. Chen, K. Takru, T. Joie Hadi Nata, and M. Li, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images," in *Proc. 7th Int. Conf. Document Analysis and Recognition,* 2003, pp. 859-864.

[5] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. of Electronic Imaging,* vol. 13, pp. 146-168, 2004.

[6] W. Niblack, An introduction to digital image processing: Prentice Hall, 1986.

[7] N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Trans. Systems Man Cybernet,* vol. 9, pp. 62-66, 1979.

[8] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *Pattern Recognition,* vol. 33, pp. 225-236, 2000.

[9] M. Sezgin and B. Sankur, "Selection of thresholding methods for nondestructive testing applications," in *Proc. 2001 Int. Conf. Image Processing,* 2001, pp. 764-767 vol.3.

[10] O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, pp. 1191-1201, December 12, 1995.

[11] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. New Jersey: Prentice-Hall, 2002.

[12] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Graph. Models Image Process,* vol. 29, pp. 273-285, 1985.

[13] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition,* vol. 19, pp. 41-47, 1986.

[14] A. Rosenfeld and P. D. l. Torre, "Histogram concavity analysis as an aid in threshold selection," *IEEE Trans. Syst. Man Cybern.,* vol. SMC-13, pp. 231-235, March-April, 1983.

[15] W. H. Tsai, "Moment-preserving thresholding: A new approch," *Graph. Models Image Process,* vol. 19, pp. 377-379, 1985.

[16] N. R. Pal and S. K. Pal, "Entropic thresholding," *Signal Process,* vol. 16, pp. 97-108, 1989.

[17] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition,* vol. 29, pp. 1335-1346, 1996.

[18] P. K. Sahoo, S. Soltani, A. K. C. Wong, and Y. C. Chen, "A survey of thresholding techniques." vol. 41: Academic Press Professional, Inc., 1988, pp. 233-260.

[19] P. Stathis, E. Kavallieratou, and N. Papamarkos, "An evaluation survey of binarization algorithms on historical documents," in *19th Int. Conf. on Pattern Recognition, ICPR 2008.,* 2008, pp. 1-4.

[20] E. Badekas and N. Papamarkos, "Optimal combination of document binarization techniques using a self-organizing map

neural network," *Engineering Applications of Artificial Intelligence,* vol. 20, pp. 11-24, February, 2007.

[21] Y. Yitzhaky and E. Peli, "A method for objective edge detection evaluation and detector parameter selection," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 25, pp. 1027-1033, 2003.