A Thresholding Algorithm for Text/Background Segmentation in Degraded Handwritten Jawi Documents

Tengku Mohd Afendi Zulcaffle, Al-Khalid Othman, Wan Azlan Wan Zainal Abidin, Shahrol Mohammaddan,
Ade Syaheda Wani Marzuki
Faculty of Engineering,
UNIMAS
Kota Samarahan, Sarawak, Malaysia
ztmafendi@feng.unimas.my

***Abstract*—The old documents in Jawi script are being used widely for references. The hard copies of those scripts will deteriorate as time passes. Most of the previous works on Jawi documents focused on the character recognition and the accuracy of the algorithm was very much affected by noise. An effective preprocessing method is required to binarize degraded Jawi document. In this paper, a new technique to threshold degraded Jawi document is proposed. The results of the new algorithm were also evaluated and compared with several algorithms. The quality of the thresholding methods was assessed using visual inspection and mathematical evaluation. The results show that the new technique has outclassed other binarization algorithms.**

***Keywords*—image processing, document binarization, degraded Jawi documents**

I. INTRODUCTION

The Jawi-script is an art of writing that has existed for centuries in South East Asia. The Jawi script is originated from Arabic script and has been adapted to suit the Malay Language. The Malay Manuscript Centre of Malaysian National Library has managed to gather more than 2,000 old Jawi manuscripts [1]. Old Jawi documents are also available in libraries of other countries like Indonesia, Brunei, South Africa, Singapore, Germany, Netherlands, USA, and France.

Several Jawi character recognition and image analysis algorithms have been developed. Zaidi et al. [2] and Zaidi et al. [3] have developed Jawi recognition algorithms. In both papers, thinning algorithms were used to binarize the Jawi characters and the algorithms failed to binarized the Jawi characters due to present of noise. Khairuddin et al. [4] conducted studies on skew detection and correction of Jawi images and utilized the Sobel edge operator to binarize the non degraded document. The Sobel edge operator is not suitable for the degraded document because it will detect the uneven illumination or noise on the background as foreground pixels. Zaidi et. al. [5] and Zaidi et. al. [6] have conducted research on segmenting long connected Jawi characters into isolated letters. In their studies, the Jawi documents were manually cropped and a single threshold value (global thresholding method) was utilized to binarize the cropped documents. Sitti Rachmawati et al. [7] suggested Sauvola and Pietikainen [14] thresholding method to binarize degraded Jawi documents.

Since the previous thresholding and binarization algorithms for Jawi documents were not studied thoroughly, the algorithms to threshold Roman documents are also reviewed. Trier and Taxt [8] have evaluated 4 global methods, 11 local methods, and 4 modified local methods. In this evaluation the Niblack's method [9] that enhanced by the Yanowitz and Bruckstein's [10] post processing step was the best overall method. In the study made by Leedham et al. [11], several algorithms were compared and the best method for historical images was the local Quadratic Integral Ratio (QIR) method [12]. The QIR had been compared with several methods in Leedham et al. [13]. However, in this evaluation the best method for historical images was the proposed Background Subtraction (BS) technique.From the evaluations made in Leedham et al. [11] and Leedham et al. [13], we can conclude that the BS technique was the best method to binarize historical documents.

The BS method and the best method in Trier and Taxt [8] evaluation, the modified Niblack's method are evaluated in this paper. Two other methods that were presented in Sauvola and Pietikainen [14] and Yanni and Horne [15] are also included in our study. These two methods were not included in the comparison in Trier and Taxt [8], Leedham et al. [11], and Leedham et al. [13]. Due to the wide usage of the Niblack's method [9] in many binarization algorithm, the method is also evaluated. Those methods were evaluated and their results were compared with our proposed technique.

The paper is organized as follows. Section II presents the proposed method. Experimental results are presented in Section III. Finally conclusion is drawn in Section IV.

II. THE PROPOSED METHOD

The proposed new binarization method separates the document into three different categories: document background, text region's background, and characters. First, the algorithm separates the document into document background and text region. The algorithm segments the characters from their backgrounds through several different levels of

978-0-7695-4269-0/10 $26.00 © 2010 IEEE
DOI 10.1109/ACT.2010.43