Bleed Through Effect Removal from Document Images

Shankha De, Arpana Rawal
shankhada2009@gmail.com
arpana.rawal@gmail.com
Bhilai Institute of Technology - Durg

Abstract:

Bleed through is the most common type of degradation effect found in document image scans particularly It is mostly seen in handwritten/printed documents that are scripted on both sides where the content of the back side appears in the front side as interference. Bleed through effect impacts the performance of optical character recognition. In this paper a simple content based global binarization method is proposed to remove bleed through effect.

Keywords: Document Image, Bleed through, Binarization, Otsu's Threshold, Opacity.

Introduction

Opacity is a measurement that describes the quality of the paper. It measures how much light can pass through the paper. The ink also passes through the paper in the other side. This ink diffusion degrades the content written on that side and so, content of the one side can be visible from the opposite site. It depends on the quality of the paper and ink used for writing [1]. This causes poor content clarity in optical character recognition process. This effect is called Bleed-through/show through effect/ back-to-front interference. Bleed through is internal form of degradation in the document image. Although , human eyes can distinguish the main content and the degraded part but optical character recognition yields the degraded response [2].

Bleed through removal methods: an overview

Bleed through removal is an open challenge as no method is found appropriately generic , handling all kinds of show-through images. Fadoua et al (2006) proposed an recursive unsupervised classification technique using principle component analysis and k-means algorithm[3]. Naga Sudha D (2015) illustrates phase base binarization for degraded document image by taking both side scan document [4]. Sharma Avinash et al (2011) used to remove bleed through effect based on Otsu's method[5]. Estrada, R., & Tomasi, C. (2009) employ hysteresis approach in two steps by threshold selection and ink growth. [6]. Ekta Vats et al,(2017) use Bayesian optimization on hyperparameter for threshold selection selection[7]. In this paper binarization is done based on the document content , its histogram features and

Otsu's multiple threshold value. Figure 1 shows the images with bleed through effect. Samples are taken from DIBCO data set [8].

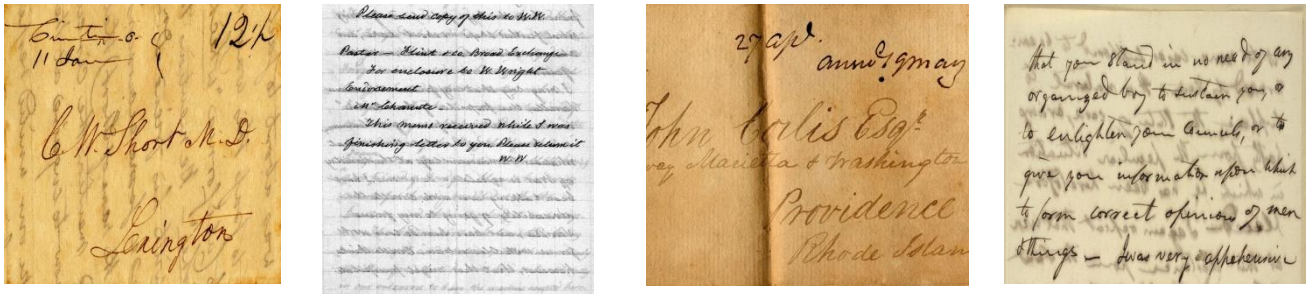


Figure 1: Input document image scans for Bleed through removal

Problem Formulation

Binarization is one the popular solution for bleed through effect. In an attempt to use Otsu's method ,as one of the most popular global binarization technique for bleed through effect removal , as it considers the interference as content. The binarization method in this paper , is proposed to be modified by histogram and Otsu's method. Histogram of the image gives the frequency of each pixel of the document image. Bleed through effect generates intermediate gray level values that lie in between foreground and background pixel values. Low contrast image makes the show through situation even more adverse. So any document image degraded by bleed through has three set of pixels classification problem. Thus, the proposed problem formulation aims to remove these intermediate pixels thereby reduce the bleed through effect.

Proposed Method

Any text document image consists of foreground (text part) that is represented by black and background part represented by white [9]. It has been observed that main content is near to black region (foreground) and occupies far less than the white (relatively bright) background region. Then intermediate gray value is the cause of bleed through. The histogram peaks pertaining to three regions: foreground, intermediate and background, shall be considered along with Otsu's multiple threshold values to compute expected gap parameter and final threshold for binarization [10].This final threshold will be used for binarization step.

The basic steps of the bleed through removal can be outlined as follows:

Step 1: Get the Input Image

Step 2: Convert it into gray scale image

Step 3: Calculate Histogram and Otsu's multiple thresholds

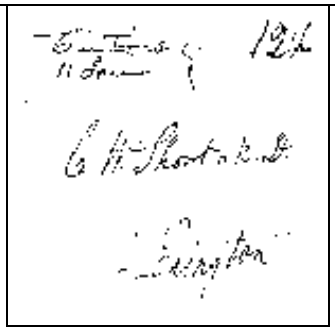
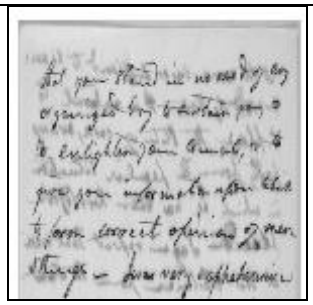
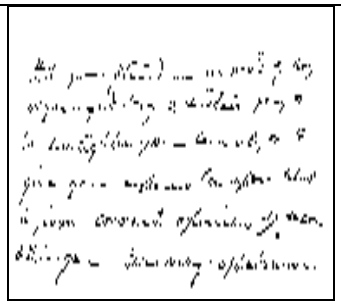
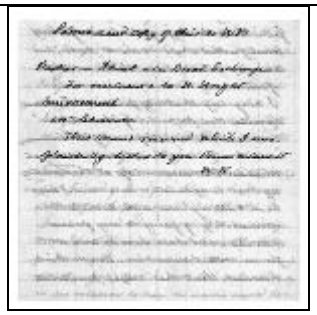
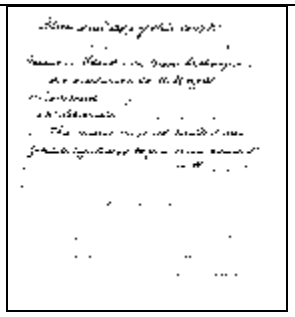
Step 4: Calculate expected gap between foregrounds and bleed through pixel.

Step 5: Calculate Final threshold value using initial, intermediate threshold and expected gap.

Step 6: Do binarization using calculated threshold value.

Experimental Result

The pilot study considered only four image scans for initiating the experiment. The images were resized to 256x256 to reduce the time and space complexity. Table 1 illustrates the experimental result on these sample images using MATLAB R2013a.

Gray Scale Image	Output Image
	
	
	

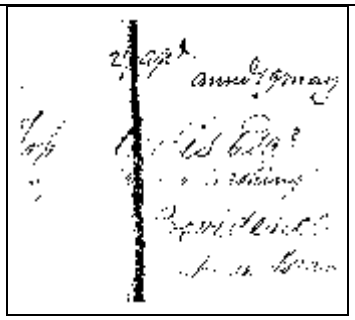
	

Table 1: Output document images after Bleed through removal

Proposed Outcome

It has been observed that output image shows only main content of the document. This method is simple and robust, although raises some remarks. If the ink colour changes in the foreground content then it will not work properly as it will not be able to distinguish different colour for foreground as per assumption. Bleed through effects removal dependence on the threshold values and gap selection between bleed thorough and foreground pixels, still persists as a scope of work, requiring further research experiments.

Acknowledgement

The paper is an experiment based initial work as an outcome carried out as a part of doctoral course curriculum, supported by research and development cell, Department of computer Science and Engineering, Bhilai Institute of Technology, Durg, Chhattisgarh, India.

References

[1] Doermann, D., & Tombre, K. (2014). *Handbook of Document Image Processing and Recognition.* Springer Publishing Company, Incorporated.

[2] Lins, R. (2009). A taxonomy for noise in images of paper document - The physical noise. *ICDAR 2009. 5627*, pp. 844-854. Berlin, Heidelberg: Springer.

[3] Drira, Fadoua & Lebourgeois, Frank & Emptoz, Hubert. (2006). Restoring Ink Bleed-Through Degraded Document Images Using a Recursive Unsupervised Classification Technique. 3872. 38-49. 10.1007/11669487_4.

[4] Naga Sudha D, Y Madhavee Latha, L Pratap Reddy,(2015) Improved Degraded Document images Using Phase Based Binarization, International Journal of Recent

Advances in Engineering & Technology (IJRAET), Volume-3, Issue -9, ISSN (Online): 2347 – 2812.

[5] Sharma, Avinash & Mahaldar, Sahil & Banerjee, Serene. (2011). Enhanced Bleed Through Removal for Scanned Document Images. Proc SPIE. 7870. . 10.1117/12.876636.

[6] Estrada, R., & Tomasi, C. (2009, July). Manuscript bleed-through removal via hysteresis thresholding. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on* (pp. 753-757). IEEE.

[7] Vats, E., Hast, A., & Singh, P. (2017, November). Automatic document image binarization using Bayesian optimization. In *Proceedings of the 2017 Workshop on Historical Document Imaging and Processing. ACM (2017, in press)*.

[8] http://vc.ee.duth.gr/h-dibco2016/ - searched on 25/01/2018.

[9] Gonzalez, R. C., & Woods, R. E. (2006). *Digital Image Processing* (3rd Edition ed.). Prentice Hall.

[10] Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE , 9* (1), 62-66.