

# Historical Document Image Enhancement Using Background Light Intensity Normalization

Zhixin Shi and Venu Govindaraju

Center of Excellence for Document Analysis and Recognition (CEDAR)

State University of New York at Buffalo, Amherst, USA

zshi@cedar.buffalo.edu

## Abstract

*This paper presents a new background light intensity normalization algorithm suitable for historical document images. The algorithm uses an adaptive linear function to approximate the uneven background due to the uneven surface of the document paper, aged color and light source of the cameras for image lifting. Our algorithm adaptively captures the background with a “best fit” linear function and normalized with respect to the approximation. The technique works for both gray scale and color images with significant improvement in readability.*

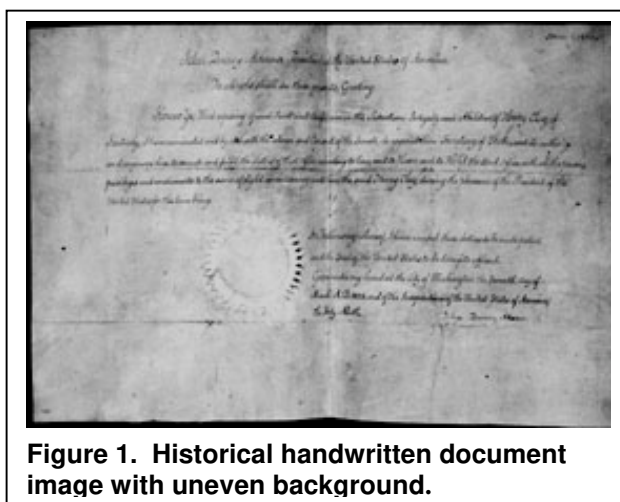
## 1. Introduction

The Library of Congress of the United States has a large collection of handwritten historical document images. The originals are carefully preserved and not easily available for public viewing. Photocopying of these documents for public access calls for enhancement methods to improve their readability for various purposes such as visual inspections by historians and OCR.

There are two types of deficiencies in the quality of historical document images. First, the original paper document is aged leading to deterioration of the paper media, ink seeping and smearing, damages and added dirt. The second problem is introduced during conversion of the documents to their digital image form. In order to best preserve the fragile originals, the digital images are usually captured by using digital cameras instead of platen scanners. The paper documents cannot be forced flat and the light source for digital cameras is usually uneven.

Previous image enhancement algorithms for historical documents have been designed primarily for segmentation of the textual content from the background of the images. An overview of the traditional thresholding algorithms for text segmentation are given in [1] which compares three popular methods, namely Otsu’s thresholding technique, entropy techniques proposed by Kapur et al. and the minimal error technique by Kittler and Illingworth. Another entropy-based method specifically designed for

historical document segmentation [2] deals with the noise inherent in the paper especially in documents written on both sides. Tan et al. presented methods to separate text from background noise and bleed-through text (from the backside of the paper) using direct image matching [3] and directional wavelets [4]. These techniques are designed mainly as preparation stages for subsequent OCR processing. Other methods for historical image enhancement are driven by improving human readability while maintaining the original “look and feel” of the documents [5].



**Figure 1. Historical handwritten document image with uneven background.**

In this paper we propose a novel technique for historical handwritten document image enhancement. Our method is targeted towards enhancing images with uneven background (Figure 1). A linear model is used adaptively to approximate the paper background. Then the document image is transformed according to the approximation to a normalized image that shows the foreground text on a relatively even background. The method works for gray scale images as well as color images.

In section 2 we present our approximation and normalization algorithm. In section 3 we describe how to use our method on color images. We present experimental results in section 4 and conclusions in section 5.

## 2. Background normalization algorithm

Global thresholding methods find a single cut-off level at which pixels in a gray scale image can be separated into two groups, one for foreground text and the other for the background. For complex document images, it is difficult to find such a global threshold. Therefore thresholding techniques use adaptive approaches that find multiple threshold levels each for a local region of an image have been proposed. The assumption made is that there exists relatively flat background at least in small local regions. This assumption does not hold well for historical document images.

Let us treat an image as a three-dimensional object whose positional coordinates are in the  $x$ - $y$  plane and the pixel gray values are in the  $z$  plane.

Consider the extreme case when a document image does not include any textual content. Then the image merely represents the paper surface. The traditional thresholding applied to this image should find a plane  $H$  parallel to  $x$ - $y$  plane above the paper surface. The plane  $H$  should also be very close to the paper surface if it is sufficiently flat.

In the case of historical documents with uneven background, although we cannot easily find such a plane  $H$  strictly above the background surface, with a carefully chosen threshold we can find a plane  $K$  above most of the background pixels with almost all of the foreground pixels above the plane. We accomplish this using a direct histogram approach. The objective is to keep most of the foreground pixels above  $K$ . The pixels below  $K$  together represent an incomplete background. Figure 2 shows this idea of selecting the threshold using the histogram curve. Figure 3 is an image of the incomplete background, in which the missing pixels are filled in white.

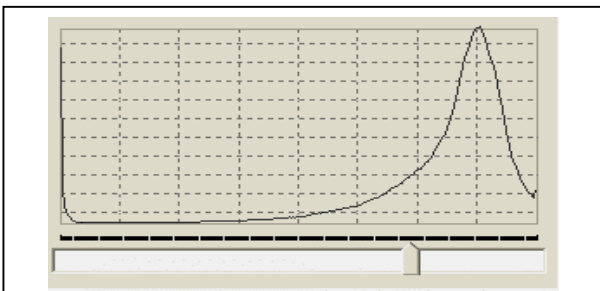


Figure 2. Histogram curve showing the peak for the back ground pixels.

### 2.1. Linear approximation

The missing background pixels could be filled by using a polynomial spline, which would find a curved surface

that fits all of the background pixels in Figure 3. But since even the leftover background in Figure 3 may still have some “high” (raised from the flat surface) pixels which are likely to be part of the foreground text. We propose a linear model using linear functions to approximate the background in small local regions. An image is first partitioned into  $m$  by  $n$  smaller regions each approximating a flat surface. In each such region we find a linear function in the form

$$Ax + By - z + D = 0 \quad (1)$$

Pixels in the leftover background are represented as points in the form  $(x_i, y_i, z_i)$  where  $(x_i, y_i)$  is the position of a pixel and  $z_i$  is the pixel value. We apply the minimal sum of distances,

$$\min \sum_i (Ax_i + By_i - z_i + D)^2 \quad (2)$$

where the sum is taken for all the available points  $(x_i, y_i, z_i)$  in the leftover background image. The minimization gives a “best fit” linear plane (1) because the distance from any point  $(x_i, y_i, z_i)$  to the plane in (1) is a constant proportional to  $|Ax_i + By_i - z_i + D|$ .

The solution for  $A$ ,  $B$  and  $D$  is obtained by solving a system of linear equations that are derived by taking the first derivatives of the sum function in (2) with respect to the coefficients, and setting the derivative functions to zeros. Therefore in each small partition we find a plane that is a best fit to the image background in the partition. The pixel value of the plane is evaluated by

$$z = Ax + By + D \quad (3)$$

for each pixel located at  $(x, y)$ .

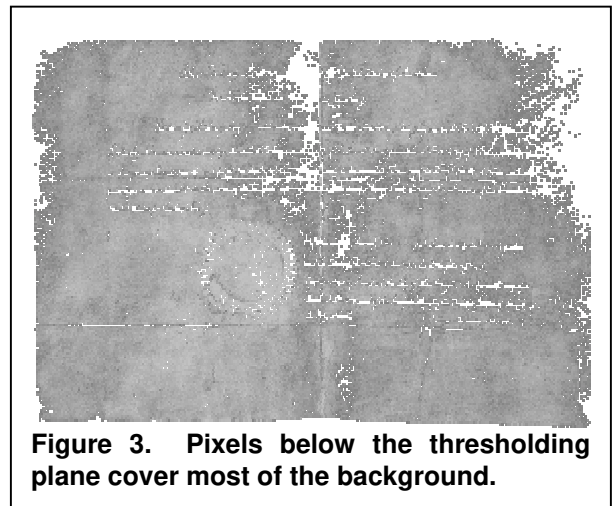


Figure 3. Pixels below the thresholding plane cover most of the background.

## 2.2. Image normalization

The original gray scale image can be normalized by using the linear approximation. Assume a gray scale image with pixel values in the range 0 to 255 (0 for black and 255 for white). For any pixel at location  $(x, y)$  with pixel value  $z_{orig}$ , the normalized pixel value is then computed as

$$z_{new} = z_{orig} - z + c \quad (4)$$

where  $z$  is the corresponding pixel value on the approximated plane in (3);  $c$  is a constant fixed to some number close to the white color value 255. Example of a normalized image is shown in Figure 4.



**Figure 4. Normalized historical document image showing an even background.**

## 3. Normalization for color images

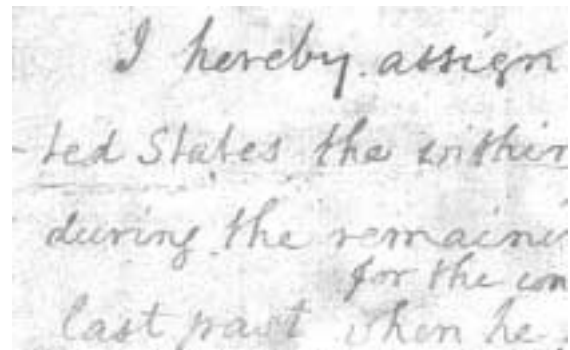
Compared to color images, gray scale images capture mainly the light intensity aspect of the original document. The uneven backgrounds of the document paper shown in the images are represented as uneven light intensity levels. Therefore our normalization algorithm also works for images with uneven background caused by uneven camera light source and discolorations. Figure 5 is an example for a damaged and discolored image.

To use the normalization algorithm on a color image, we first apply a color system transformation to change the color image in its RGB representation to its representation in YIQ color system by the following transform (5).

$$\begin{cases} Y = 0.2992R + 0.5868G + 0.1140B \\ I = 0.5960R - 0.2742G - 0.3219B \\ Q = 0.2109R - 0.5229G + 0.3120B \end{cases} \quad (5)$$

The YIQ system is the color primary system adopted by National Television System Committee (NTSC) for color TV broadcasting. Its purpose is to improve transmission efficiency and for downward compatibility with black and white television. The human visual system is more sensitive to changes in luminance than to changes in hue or saturation, and thus a wider bandwidth should be dedicated to luminance than to color information.  $Y$  is similar to perceived luminance;  $I$  and  $Q$  carry color information and some luminance information. Since the YIQ system is more sensitive to change in luminosity, the light intensity variation due to the uneven background of historical document images is mostly captured in the  $Y$  channel.

As we described in section 2, we apply our background normalization algorithm on the split  $Y$ -image. The YIQ design gives both  $I$  and  $Q$  channels very narrow



**Figure 5. A very difficult example: Image of the sale contract between Thomas Jefferson and James Madison. Above is the original image and below is a portion of the normalized image.**

bandwidth. For the purpose of background normalization we take a single value for  $I$  and  $Q$  in each small partition, respectively. The single values are calculated by averaging the corresponding pixel values in the partition.

To get back to RGB color system we use the YIQ to RGB transform that is the inverse transform of (5). The normalized  $Y$  channel together with the locally averaged  $I$  and  $Q$  channels are transformed back to a RGB color image to yield the normalized document image.

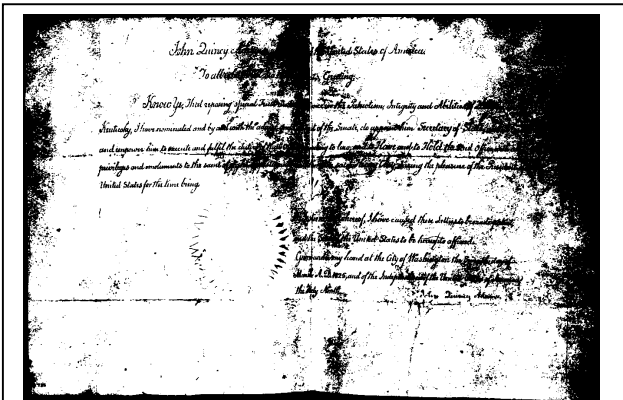


Figure 6. It is impossible for the original document image in Figure 1 to be segmented using a global threshold. The best global threshold we could manually find is at 200 and the binarized image shows significant parts of text obliterated.

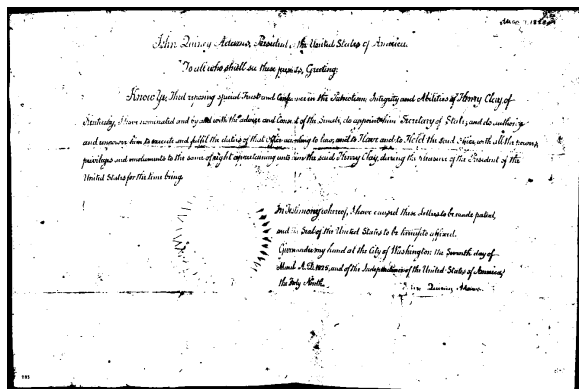


Figure 7. The image normalized using the method described in this paper can be easily binarized using a global threshold.

## 4. Experiment and results

We have downloaded 100 historical handwritten document images from the Library of Congress website. These images are selected because they all have obvious uneven background problems. Visual inspections of the enhanced images show a marked improvement in image quality for human reading.

Furthermore, we have chosen 20 images from the set. These are images impossible to be segmented by any global threshold value. Our method successfully finds a better binarized image; see example images in Figure 6 and 7.

## 5. Conclusions

In this paper we present a historical handwritten document image enhancement algorithm. The algorithm uses a linear approximation to estimate the “flatness” of the background. The document image is normalized by adjusting the pixel values relative to the line plane approximation. From our experiments and visual evaluation, the algorithm has been found to work successfully in improving readability of document images on aged paper, wrinkled paper and camera images with uneven light sources.

## References

- [1] G. Leedham, S. Varma, A. Patankar, V. Govindaraju “Separating Text and Background in Degraded Document Images - A Comparison of Global Thresholding Techniques for Multi-Stage Thresholding” Proceedings Eighth International Workshop on Frontiers of Handwriting Recognition, September, 2002, pp. 244-249
- [2] C.A.B.Mello and R.D.Lins, “Image Segmentation of Historical Documents”, Visual2000, Mexico City, Mexico, September 2000.
- [3] Q. Wang and C.L. Tan, “Matching of double-sided document images to remove interference”, IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA, 2001.
- [4] W. Wang, T. Xia, L. Li and C.L. Tan, “Document image enhancement using directional wavelet”, Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, June 18 - 20, 2003
- [5] C.A.B.Mello and R.D.Lins, “Generation of Images of Historical Documents by Composition”, ACM Symposium on Document Engineering, McLean, VA, USA, 2002.