Binarization of Camera-Captured Document using A MAP Approach

Xujun Peng, Srirangaraj Setlur, Venu Govindaraju and Ramachandrula Sitaram†
CUBS, University at Buffalo, Amherst, NY 14228, USA
†HP Labs India, Bangalore 560030, India

ABSTRACT
Document binarization is one of the initial and critical steps for many document analysis systems. Nowadays, with the success and popularity of hand-held devices, large efforts are motivated to convert documents into digital format by using hand-held cameras. In this paper, we propose a Bayesian based maximum a posteriori (MAP) estimation algorithm to binarize the camera-captured document images. A novel adaptive segmentation surface estimation and normalization method is proposed as the preprocessing step in our work and followed by a Markov Random Field based refine procedure to remove noises and smooth binarized result. Experimental results show that our method has better performance than other algorithms on bad or uneven illumination document images.

Keywords: Document Binarization, Markov Random Field, Image Processing, Camera-captured

1. INTRODUCTION
Because most document retrieval and recognition methods rely on high quality binarized document images, the document image binarization which segments foreground text area from blank background plays a key role in document analysis and recognition system. With the vast amount of using of the hand-held devices, more and more documents can be captured and converted into digital format using hand-held devices such as PDA and hand-held camera with high-resolution. However, despite the convenience of the hand-held devices, the camera-captured document images easily suffer the degradation such as distortion, uneven or bad illumination, etc which cause the binarization be a challenge problem.

The research into the image binarization can be traced back to the pioneering work on global threshold based methods, such as the famous Otsu’s method which used a single threshold value that maximizes the inter-class (foreground and background) variance or minimizes the intra-class variance to segment the entire image. The drawback of this method is that it assumes the histogram of images has two distinct peaks for different classes respectively and can be separated. But this assumption is hardly satisfied in most real applications, especially to camera-captured document images. Fig. 1 shows examples of degraded camera-captured document images, along with their corresponding histograms.

To overcome the disadvantage of single global threshold binarization method, Valizadeh et al. suggested mapping the original grey level of each pixel to a new domain prior to using global threshold. Shi and Govindaraju normalized grey level of pixels within degraded historical document image and binarized them through the global thresholding. By using a two rounds polynomial surface smoothing process, Lu and Tan used similar manner of normalization before thresholding.

Another thrust has been on using the adaptive threshold or local threshold which computes the threshold of each pixel according to the properties of its own or its neighbor pixels. Niblack proposed a method which uses mean value and variance of small window to determine the threshold of centered pixels. The potential problem of this method is that large amount of noises are produced in pure blank background areas and it is sensitive to the window size. An extension of Niblack’s method was described by Sauvola et al. which obtained better performance in open background regions. In Gatos’s work, a foreground and background surface estimation algorithm was proposed followed by an adaptive thresholding procedure. Instead of using Otsu threshold globally,
We can eliminate the third term log \( Pr(Y) \) of Eq.1 since it is not dependent on \( X \). The MAP estimation of

\[
\hat{X} = \arg \max_{\hat{X}} Pr(\hat{X}|Y) = \arg \max_{\hat{X}} \frac{Pr(Y|X) Pr(X)}{Pr(Y)}
\]

\[= \arg \max_{\hat{X}} \{ \log Pr(Y|X) + \log Pr(X) - \log Pr(Y) \} \] (1)

We can eliminate the third term log \( Pr(Y) \) of Eq.1 since it is not dependent on \( X \). The MAP estimation of

Moghaddam and Cheriet suggested an adaptive Otsu method in local area. In their work, the stroke width and line height were used as prior knowledge which makes the binarization result not to be sensitive to the window size. By dividing entire camera-captured document image into several small sized pieces, Chou et al. used a SVM based algorithm to classify each small area as one of four classes and take different binarization actions accordingly.

In recent years, Markov Random Field (MRF) based image restoration algorithms have gained great interests from researchers in document processing areas. Considering the ideal document image as a black-white image which is down-sampled and blurred to grey image by adding Gaussian noises, document binarization can be looked at as a special restoration problem. Lelore and Bouchara proposed a MRF model for binarization which can remove the noise and improve the character connectivity. Lettner et al. used a similar framework but defined the Gibbs distribution in a different form to binarize the degraded documents. To binarize uneven lighted document, Kuk and Cho initially segmented entire document image using mean filter responses as features and relabeled the pixels using a minimization technique which can be looked at as a variant of Gibbs model.

In this paper, we describe a MAP based algorithm to binarize camera-captured document images with bad or uneven illumination. The overview of MAP based statistic model for binarization is presented in section 2. In section 3, we propose a novel preprocessing method to estimate an adaptive segmentation surface and normalize entire document image based on it. A MRF based refine procedure is introduced in section 4. Experimental results and conclusions are covered in section 5 and section 6.

2. STATISTIC MODEL

The document image binarization which assigns each pixel of the document as one of two labels (foreground or background) can be modeled as an restoration process to estimate a binarized ideal image \( X \) given only the blurred degraded camera-captured image \( Y \) with uneven or bad illuminations. By taking Bayesian rule and using log function, the maximum a posteriori (MAP) estimation of \( X \) can be applied:

\[
\hat{X} = \arg \max_{\hat{X}} Pr(\hat{X}|Y) = \arg \max_{\hat{X}} \frac{Pr(Y|X) Pr(X)}{Pr(Y)}
\]

\[= \arg \max_{\hat{X}} \{ \log Pr(Y|X) + \log Pr(X) - \log Pr(Y) \} \] (1)

We can eliminate the third term log \( Pr(Y) \) of Eq.1 since it is not dependent on \( X \). The MAP estimation of
\( \mathcal{X} \) can then be rewritten as:

\[
\hat{X} = \arg \max_{\mathcal{X}} \{ \log Pr(Y|\mathcal{X}) + \log Pr(\mathcal{X}) \}
\]

\[
= \arg \min_{\mathcal{X}} \{ -\log Pr(Y|\mathcal{X}) - \log Pr(\mathcal{X}) \}
\]

Commonly, the optimization of Eq.2 can be achieved by minimizing an energy function: \(^{13,14}\)

\[
E(\mathcal{X}) = \sum_{i \in \mathcal{V}} U_i(x_i) + \sum_{(i,j) \in \mathcal{E}} V_{i,j}(x_i, x_j)
\]

where \( \mathcal{X} = \{x_i | i \in \mathcal{V}\} \) is a labeling of a lattice which corresponds to the document image. \( U(x_i) \) is an unary penalty function and is derived from log-likelihood \( \log Pr(y_i|x_i) \) which denotes the probability of the label of site \( i \) given its observation \( y_i \). \( \mathcal{E} \) is a set of edges for \( \mathcal{Y} \) which connect neighboring pixels in a 4-neighbors lattice connectivity. \( V_{i,j}(x_i, x_j) \) is a pairwise potential function which is derived from log-prior \( \log Pr(x_i) \). Normally, \( V_{i,j}(x_i, x_j) \) which indicates the prior knowledge of document image has a Gibbs distribution and encourages global smoothness of entire image.

### 3. PRE-PROCESSING

Due to uneven or bad illumination, the intensity of background is not consistent within the camera-captured document image which causes the single threshold based method impractical. We estimate an initial adaptive segmentation surface using local information around every single pixel and the segmentation surface is used to normalize the entire document.

Prior to the estimation of adaptive segmentation surface, a \( m \times n \) sized window is centered on every pixel \( i \) to compute corresponding mean value \( \mu_i \) and variance \( \delta_i \). The maximum variance \( \delta_{\text{max}} \) and minimum variance \( \delta_{\text{min}} \) over entire image are also obtained. Normally, for a given pixel \( i \), the variance is large if it belongs to the text area and the variance is small if the pixel comes from pure background areas. For this reason, we propose an adaptive threshold for every pixel using logistic function as described in Eq. 4 where the threshold has a smaller value than corresponding mean value \( \mu_i \) if its variance \( \delta_i \) is small and has a value close to mean value if the variance of this pixel is large:

\[
S(i) = \mu_i \begin{cases} 
1 - k & \text{if } \delta_i \geq \delta_{\text{min}} \\
1 + e^{-B\left(\frac{\delta_i - \delta_{\text{max}}}{\delta_{\text{max}} - \delta_{\text{min}}} - M\right)}^{1/\nu} + k & \text{if } \delta_i < \delta_{\text{min}} \end{cases}
\]

where \( B \) controls the growth rate of the logistic curve, \( M \) and \( \nu \) affect the time of maximum growth occurs and \( k \) is the minimum value the curve can achieve. The final parameters we chose in our experiment are \( B = 25 \), \( M = 0.005 \), \( \nu = 20 \) and \( k = 0.97 \).

To enlarge the difference between foreground and background, we map the original document image into a new domain using Eq. 5:

\[
F(i) = \begin{cases} 
128 + 128 \times \left( \frac{I(i) - S(i)}{\delta_{\text{max}}} \right)^{\alpha} & \text{if } \delta_{\text{max}} \geq S(i) \\
128 - 128 \times \left( \frac{S(i) - I(i)}{\delta_{\text{max}}} \right)^{\alpha} & \text{if } \delta_{\text{max}} < S(i) 
\end{cases}
\]

where \( S(i) \) is the adaptive threshold obtained from Eq. 4 for pixel \( i \), \( I(i) \) is the original gray intensity of the pixel, \( df_{\text{max}} \) is the maximum difference between foreground and segmentation surface and \( db_{\text{max}} \) is the maximum difference between background and segmentation surface respectively, \( \alpha \) is a parameter which controls the growth rate of mapping curve and is set to be 0.1 in our experiment.

Fig. 2 shows the preprocessing result of a partial hand-held device captured document image which has bad illumination as shown in Fig. 1(a). Fig. 2(a) illustrates the corresponding adaptive segmentation surface for this image and Fig. 2(b) is the normalized result. The normalized image is used as the observation of our Markov Random Field (MRF) based binarization framework which is described in the following section.
4. ENERGY FUNCTION

In our proposed binarization framework, the document image is initially binarized using adaptive segmentation surface and followed by a MRF based refine procedure. As described in section 2, the binarization can be considered as a MAP procedure which utilizes the observation of the image and pairwise relationship between pixels that are formulized as unary energy function and pairwise energy function respectively. The merit of MRF based binarization is that it not only takes the property of each single pixel into consideration, but uses the surrounding pixels property to refine the binarization result. Unlike other MRF based algorithms\textsuperscript{10–12} only consider the global smoothness of the document image, we propose a novel stroke based feature and energy criteria which preserve the global smoothness and the edge of strokes at the same time.

4.1 Stroke Features

Besides intensity feature which is derived from the normalized image, a stroke related feature is carried out in this paper. Prior to feature extraction, the document image is initially binarized using a single global threshold on the normalized image. We calculate the shortest length from every foreground pixel to background and record the maximum length within each connected component. The shortest length for each pixel is revised by subtracting this value from the maximum length within the connected component. The new feature measures the distance from a foreground pixel to the inner center of the corresponding connected component or the potential to be on the edge and is denoted as $s(i)$ for pixel $i$. Fig. 3 shows an example of character $n$ and its corresponding stroke feature. The average stroke width for entire document image is also computed and denoted as $sw$.

4.2 Unary and Pairwise Energy Function

The goal of MRF based refine procedure is to remove noises and smooth the entire document image. As described in section 2, the log-likelihood $\log Pr(y_i|x_i)$ can be approximately represented by an unary energy function $U_i(x_i)$ which forces the label $x_i$ of pixel $i$ to be close to its observation $y_i$. We define $U_i(x_i)$ as:

$$U_i(x_i) = \lambda|y_i - x_i|$$  \hspace{1cm} (6)
where \( y_i \) is the normalized intensity of pixel \( i \) whose range is from 0 to 255, \( x_i \) is the label of the pixel which takes value 0 as foreground and 255 as background, parameter \( \lambda \) influences the weight of unary energy compared to pairwise energy introduced in Eq. 7.

In general, log-prior log \( Pr(x_i) \) or pairwise energy function \( V_{i,j}(x_i, x_j) \) in MRF framework measures the similarity between hidden neighbors and tends to encourage smoothness across the entire image. In this paper, we use the 4-neighbors connectivity and edge system \( E \) as notated in Eq. 3. To each pixel \( i \), we calculate the similarity between this pixel and its neighbor considering their intensity difference and edge potential properties and define the pairwise energy function as:

\[
V_{i,j}(x_i, x_j) = \left\{ \begin{array}{ll}
\alpha \exp \left( \frac{1}{\|s(i)-s(j)\|^2-(sw/2)^2} \right) + \beta \exp \left( \frac{|y_i-y_j|}{256} \right) & \text{if } x_i = x_j \\
\alpha \exp \left( \frac{1}{\|s(i)-s(j)\|^2-(sw/2)^2} \right) + \beta \exp \left( \frac{|y_i-y_j|}{256} \right) & \text{if } x_i \neq x_j
\end{array} \right. 
\]

(7)

where \( s(i) \) and \( s(j) \) is the stroke related feature (edge potential) for pixel \( i \) and its neighbor \( j \) obtained from section 4.1, \( sw \) is the average stroke width for entire document image, \( y_i \) and \( y_j \) are normalized intensity for pixel \( i \) and \( j \) respectively, \( \alpha \) and \( \beta \) are two parameters control the influence of edge potential and intensity differences.

The underlying idea of Eq. 7 is that if two neighboring pixels \( i \) and \( j \) are from the same source (with the same label \( x_i \) and \( x_j \)), their edge potential difference \( s(i) - s(j) \) should be close to 0 which causes the energy of the first term of the top energy function of Eq. 7 to be low and their intensity difference \( y_i - y_j \) also should be close to each other which makes the second term of the top energy function to have a small value either. Similarly, if two neighboring pixels are from different source, their edge potential difference is close to half of the average stroke width and they have larger intensity difference which lead to a low energy for the bottom energy function of Eq. 7.

The minimum energy for overall energy function \( E(\mathcal{X}) \) is achieved by using graph cut algorithm.\(^{13}\)

5. EXPERIMENTAL RESULTS AND DISCUSSION

The target of our algorithm is to enhance the quality and readability of binarized document image and facilitate the further document analysis and retrieval. In our experiments, we captured 28 pages of research paper image using a hand-held cell-phone camera with the resolution of 3.2 mega pixels in office environment. Each page of research paper is in two column style and takes over at least 95% proportion of the image. All these document images are suffered from insufficient or uneven illumination problem along with out of focus blur and distortion.

To evaluate the performance of our algorithm, we compared the proposed method with the global threshold based algorithm Otsu\(^1\) and two local threshold based methods Niblack\(^2\) and Sauvola\(^6\) visually. Table. 1 shows one of the result images using different binarization methods. The first row is the origin document image with its corresponding text. The Otsu algorithm which used a global threshold failed to segment the text from background area as shown in the second figure of the table because of the uneven illumination. Figure in the third row of the table shows the binarization result of Niblack method which introduced a lot of noises in blank background area. The binarized image of using Sauvola method and Lu’s method\(^4\) have better performance than Niblack method where less noise are presented as shown in the fourth and fifth figure. The last figure in the table shows the result of the proposed MRF based method which outperforms other three methods. From the last figure, it can be seen that noises are removed from document image and the stroke of characters are enhanced which is as supposed in Eq. 7.

Further more, we used OCR software tesseract\(^{15}\) to evaluate our approach for those image portions with bad or uneven illumination. The last column of the Table 1 shows the OCR results of different methods. Mean Levenshtein distances\(^{16}\) between OCR results and ground truth were calculated and shown in Fig. 4, from which it can be observed that the OCR result of our proposed algorithm has the shortest Levenshtein distance to the ground truth which means more reliable OCR result than other approaches.
In this work, we assume that the values of these parameters are known a priori and are constant on each image. Below, a few characteristic lengths are defined that will be used throughout this work.

### Table 1. Binarization result of uneven illuminated document image using proposed algorithm compared with other methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Binarization Result</th>
<th>OCR Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original image</td>
<td><img src="image1" alt="Image" /> In this work, we assume that the values of these parameters are known a priori and are constant on each image. Below, a few characteristic lengths are defined that will be used throughout this work.</td>
<td><img src="ocr1" alt="OCR Result" /></td>
</tr>
<tr>
<td>Otsu</td>
<td><img src="image2" alt="Image" /> In this work, we assume that the values of these parameters are known a priori and are constant on each image. Below, a few characteristic lengths are defined that will be used throughout this work.</td>
<td><img src="ocr2" alt="OCR Result" /></td>
</tr>
<tr>
<td>Niblack</td>
<td><img src="image3" alt="Image" /> In this work, we assume that the values of these parameters are known a priori and are constant on each image. Below, a few characteristic lengths are defined that will be used throughout this work.</td>
<td><img src="ocr3" alt="OCR Result" /></td>
</tr>
<tr>
<td>Sauvola</td>
<td><img src="image4" alt="Image" /> In this work, we assume that the values of these parameters are known a priori and are constant on each image. Below, a few characteristic lengths are defined that will be used throughout this work.</td>
<td><img src="ocr4" alt="OCR Result" /></td>
</tr>
<tr>
<td>Lu’s method</td>
<td><img src="image5" alt="Image" /> In this work, we assume that the values of these parameters are known a priori and are constant on each image. Below, a few characteristic lengths are defined that will be used throughout this work.</td>
<td><img src="ocr5" alt="OCR Result" /></td>
</tr>
<tr>
<td>Proposed</td>
<td><img src="image6" alt="Image" /> In this work, we assume that the values of these parameters are known a priori and are constant on each image. Below, a few characteristic lengths are defined that will be used throughout this work.</td>
<td><img src="ocr6" alt="OCR Result" /></td>
</tr>
</tbody>
</table>
6. CONCLUSIONS

In this paper, we present a binarization algorithm which is focusing on hand-held devices captured document images with insufficient or uneven illumination. The algorithm mainly contains two key parts: initial segmentation/normalization and MAP based binarization. Initially, a logistic based non-linear function is proposed to estimate the segmentation surface for entire document image and then the image is normalized by using segment surface and another non-linear function. The MAP based binarization procedure uses a MRF framework to segment the foreground text from background area. In this framework, the normalized document image is used as the observation, and a novel pairwise energy function is defined in this paper to measure the relationship between neighboring pixels which preserves the stroke edge, removes the noises and smoothes the entire document image at the same time. Experiment results show that our method outperforms other approaches.

REFERENCES


