Optik 164 (2018) 218–223

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

Original research article

Degraded historical document image binarization using local features and support vector machine (SVM)

Wei Xiong*, Jingjing Xu, Zijie Xiong, Juan Wang, Min Liu

Hubei Collaborative Innovation Center for High-Efficiency Utilization of Solar Energy, School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, 430068, China

ARTICLE INFO

Article history:
Received 27 May 2017
Accepted 20 February 2018

Keywords:
Document image binarization
Thresholding
Segmentation
Support vector machine
SVM

ABSTRACT

This paper presents a *support vector machine* (SVM) based method for degraded historical document image binarization. Given a degraded historical document image, the proposed method first segments the image into $w \times w$ regions and implements a local contrast enhancement in each image block. We then use a SVM to select an optimal global threshold for binarization of each image block. Finally, the entire image is further binarized by a locally adaptive thresholding method. The proposed method has been evaluated over the recent *Document Image Binarization Competition* (DIBCO) datasets. The experimental results show that our proposed method outperforms other state-of-the-art techniques in terms of F-measure, NRM, DRD, and MPM.

1. Introduction

Binarization (also referred to as thresholding or foreground/background segmentation) is one of the most important preprocessing steps in *document analysis and recognition* (DAR) related applications, such as video caption extraction and text information retrieval. It aims to extract the foreground text from the document background. The performance of this preprocessing step directly affects the accuracy of the subsequent tasks, e.g., page layout analysis and *optical character recognition* (OCR). Due to the presence of different types of degradations, such as page stains, ink bleed through, and contrast variations, as shown in Fig. 1, binarization for degraded historical document images is challenging.

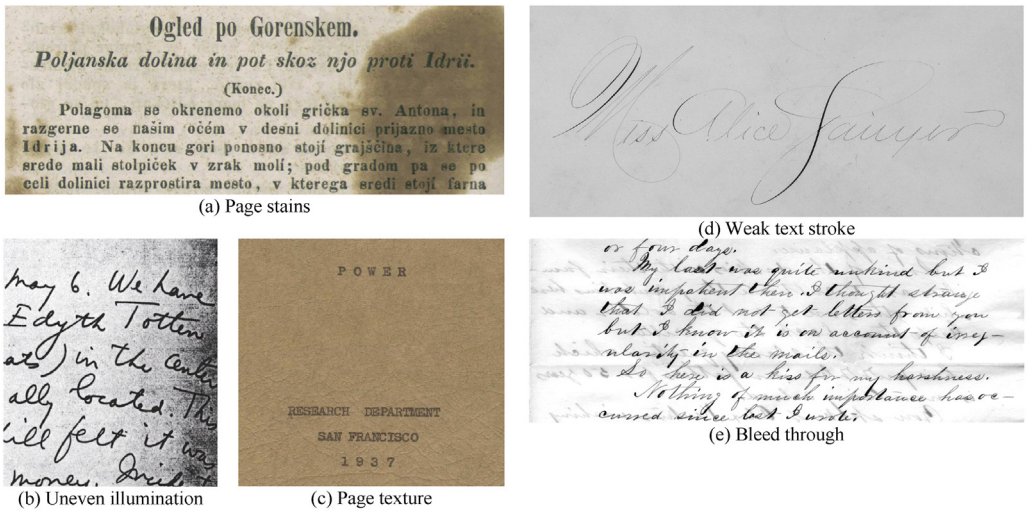
Many document image binarization techniques have been proposed and can be roughly classified as global or local [1]. Global thresholding uses a single threshold value for the entire image, for instance, the Otsu's method [2] calculates the optimum threshold to separate the foreground and the background pixels, so that their inter-class variance is maximum or equivalently their intra-class variance is minimum. The global methods exhibit relatively good performance if the histogram follows bimodal distribution, but if the image quality is too low, the binarization result is usually unacceptable.

Local thresholding uses neighborhood features to compute the threshold for each pixel. The Niblack's [3], Sauvola's [4], and Wolf's [5] methods use local average and standard deviation of the pixels within a small neighborhood window to obtain a thresholding segmentation surface. They generally have better performance but depend heavily on the sliding window size and hence the text stroke width.

Many hybrid thresholding methods have also been proposed, e.g., histogram matching methods [6,7], *convolution neural network* (CNN) methods [8], Gaussian mixture modeling methods [9], blind deconvolution methods [10], *conditional random*

* Corresponding author.
E-mail address: xw@mail.hbut.edu.cn (W. Xiong).

https://doi.org/10.1016/j.ijleo.2018.02.072
0030-4026/



(a) Page stains (b) Uneven illumination (c) Page texture (d) Weak text stroke (e) Bleed through

Fig. 1. Different types of document degradations in DIBCO testing datasets.

(a) Local image contrast (b) Traditional image gradient

Fig. 2. The local contrast and traditional gradient of the degraded historical document image in Fig. 1(a).

field (CRF) methods [11], background estimation methods [12,13], local contrast methods [14–16], Laplacian energy based methods [17,18], and multi-spectral imaging methods [19,20]. Since combined different kinds of image information, the time complexity of these hybrid approaches is relatively high.

In this paper, we propose a SVM-based binarization method that can efficiently segment the text (object) from degraded historical document background. Given a gray-scale document image, we first divide the image into $w \times w$ regions, and then utilizes a SVM to classify the document image blocks into 3 categories. Different global thresholds are used for different types of image blocks, and the document text is then roughly segmented. Finally, the text stroke width can be estimated, and thus the document text is then precisely segmented.

The major contribution is that we propose a supervised learning based technique for degraded historical document image binarization. The remainder of this paper is organized as follows: Section 2 presents the detailed methodology, Section 3 discusses the evaluation results, and Section 4 concludes the paper.

2. Methodology

Given a historical document image, we first conduct the color-to-gray conversion and then divide the image into $w \times w$ regions. Next, we enhance the local contrast and adopt the support vector machine to classify the image blocks into 3 categories. According to the classification results, an optimal global threshold is chosen to roughly separate the foreground and background. After thresholding, the image blocks are seamless spliced. The text stroke width and sliding window size can be estimated, so as to achieve a more precise segmentation.

2.1. Local contrast enhancement

We adopt Su's local contrast [16] that is defined as $C(x, y) = \frac{f_{\max}(x,y) - f_{\min}(x,y)}{f_{\max}(x,y) + f_{\min}(x,y)}$, where $f_{\max}(x, y)$ and $f_{\min}(x, y)$ represent the maximum and minimum gray-scale levels within a local neighborhood window, the size of which is 3×3 pixels in our implemented system. The denominator suppresses the background variations and makes the binarization more robust against local degradation, as depicted in Fig. 2.