Pattern Recognition 74 (2018) 225–240

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Degraded document image binarization using structural symmetry of strokes

Fuxi Jia, Cunzhao Shi*, Kun He, Chunheng Wang, Baihua Xiao

The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, University of Chinese Academy of Sciences, No.95 Zhongguancun East Road, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 19 December 2016
Revised 8 June 2017
Accepted 19 September 2017
Available online 23 September 2017

Keywords:
Document image binarization
Structural symmetry of strokes
Local threshold
Stroke width estimation

ABSTRACT

This paper presents an effective approach for the local threshold binarization of degraded document images. We utilize the structural symmetric pixels (SSPs) to calculate the local threshold in neighborhood and the voting result of multiple thresholds will determine whether one pixel belongs to the foreground or not. The SSPs are defined as the pixels around strokes whose gradient magnitudes are large enough and orientations are symmetric opposite. The compensated gradient map is used to extract the SSP so as to weaken the influence of document degradations. To extract SSP candidates with large magnitudes and distinguish the faint characters and bleed-through background, we propose an adaptive global threshold selection algorithm. To further extract pixels with opposite orientations, an iterative stroke width estimation algorithm is applied to ensure the proper size of neighborhood used in orientation judgement. At last, we present a multiple threshold vote based framework to deal with some inaccurate detections of SSP. The experimental results on seven public document image binarization datasets show that our method is accurate and robust compared with many traditional and state-of-the-art document binarization approaches based on multiple evaluation measures.

1. Introduction

Document image binarization is a fundamental step in most document analysis systems, which aims to extract text objects from the background [1,2]. The performance of subsequent steps is highly dependent on the success of binarization. But there are still many challenges when the document images contain various degradations, such as faint characters, bleed-through background, ink stains and so on. Therefore, the study on binarization for document images, in particular degraded images, is very essential.

A successful binary result preserves meaningful information while discarding noisy information [1]. It is difficult to get such a successful result when we apply the traditional local thresholding methods [3,4] to binarize the degraded document images. The reason might lie in the fact that they compute a unique threshold using all the pixels in neighborhood including the possible random noise and background disturbance. In this paper, we use multiple threshold values computed by the SSPs of the region to find out whether one pixel belongs to the foreground or not. As shown in Fig. 1, the SSPs denote the stroke edges which contain both text and background pixels. The intensity statistic of these pixels is a good approximation of the local threshold used to distinguish text from background.

The concept of SSP has first been proposed in our previous work [5]. Its effectiveness has already shown in that paper. However, since the original SSP extraction method is primitive and parameter-dependent, the binarization result is not always satisfied. In this paper, we present a modified extraction algorithm with minimum parameter tuning to extract the SSP more precisely. Specifically, the improvements are threefold. First, we use a more effective background removal process to obtain the compensated image. The gradient map is produced by the compensated one so as to deal with many types of degradations. Then we propose an adaptive threshold selection algorithm to compute a global threshold adaptively. We use this threshold to binarize the gradient magnitude map in order to extract the SSP candidates with large magnitude precisely. Finally, we adopt an iterative algorithm to estimate the text stroke widths and remove noise at the same time. By applying the orientation symmetry judgement based on the estimated stroke widths on the SSP candidates, we extract the real SSPs satisfying opposite orientation constraint in neighborhood. The other contribution of this paper is the voting framework in which we use multiple threshold values to decide whether each pixel belongs to text or not. In this way, some inaccurate detection

* Corresponding author.
E-mail addresses: cunzhao.shi@ia.ac.cn (C. Shi), chunheng.wang@ia.ac.cn (C. Wang).

https://doi.org/10.1016/j.patcog.2017.09.032
0031-3203/© 2017 Elsevier Ltd. All rights reserved.

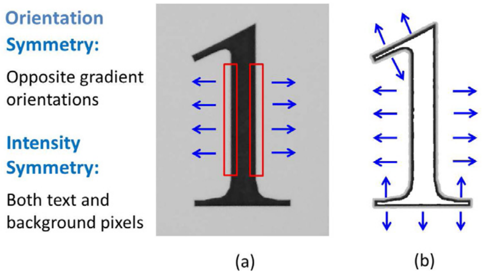
Orientation Symmetry:
Opposite gradient orientations
Intensity Symmetry:
Both text and background pixels
(a)
(b)

Fig. 1. The illustration of structural symmetric pixels (SSPs). (a) The motivation of SSP : pixels around strokes contain both text and background candidates. (b) SSP (white pixels represent Non-SSP, black and gray pixels denote text and background candidates respectively). The blue arrows denote the gradient orientations of stroke edges. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

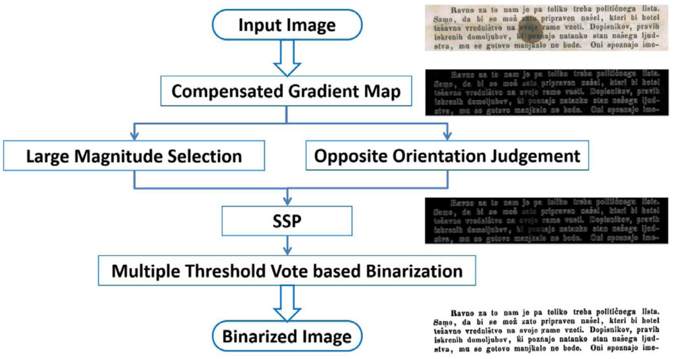
Input Image
Compensated Gradient Map
Large Magnitude Selection
Opposite Orientation Judgement
SSP
Multiple Threshold Vote based Binarization
Binarized Image

Fig. 2. The flowchart of the proposed binarization method.

of SSP can be compensated. The executable code of our algorithm can be download from the following url[1].

To demonstrate the effectiveness of the proposed method, we conduct comprehensive experiments on different datasets with various types of degradations. The experimental results show that our method achieves promising performance compared with many traditional and state-of-the-art document binarization algorithms tested on the datasets of DIBCO'09 [6], H-DIBCO'10 [7], DIBCO'11 [8], H-DIBCO'12 [9], DIBCO'13 [10], H-DIBCO'14 [11] and H-DIBCO'16 [12], based on various evaluation measures, including the F-measure, pseudo F-Measure, NRM, PSNR, DRD, and MPM.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work. In Section 3, we introduce the proposed binarization method using the structural symmetry of strokes in detail. The experimental results are presented in Section 4, and dummyTXdummy- conclusions are drawn in Section 5.

2. Related work

Document image binarization is to convert a color or gray image into a binary image, where the text and background pixels are marked in black and white respectively. Basically, the thresholding technique, which is one of the most useful binarization methods, are of three main types: global, local and hybrid [13,14]. In the global techniques, a single threshold value is calculated from whole image. The local techniques compute a local threshold based on the statistics in the neighborhood of each pixel. Hybrid method is the combination of the local and global one. For document images of good quality, global methods, like Otsu [15] and Kittler [16], is capable of extracting the text efficiently. However, for document images suffering from different types of degradations, local methods, such as Bernsen [17], Niblack [3] and Sauvola [4], usually produce better binarization results [18]. But they tend to introduce some background noise. The reason may lie in the fact that the local threshold is calculated by all pixels in neighborhood including the background disturbance.

To obtain more satisfactory binary results, Lu et al. [19] compute the local threshold only based on the stroke edge pixels and the performance is improved to some extent. To extract the stroke edges, they first estimate the background surface through an one-dimensional iterative polynomial smoothing procedure [20] and then use Otsu's method to binarize the compensated gradient map. While the gradient map is replaced with a local contrast image built with local maximum and minimum in Su's paper [21]. In [22], a more robust feature map is produced by combining the gradient map with the local contrast image. This method performs better than those in [19] and [21]. But these methods have a certain limitation as stated in [19]: the result may remain some bleed-through noise or ignore some faint characters since the final threshold is based on the local contrast. Lelore et al. [23] introduce the FAIR binarization algorithm based on a double-threshold edge detection approach and the detection strategy makes it possible to catch small details while remaining robust against noise.

The hybrid techniques combine the global and local thresholding techniques. It takes the advantages of both the techniques. Chou et al. [24] divide an input image into blocks and choose different binarization methods for each block. In [25], several methods are combined based on a vote on their outputs. In [26], the background is estimated first by performing inpainting, then a combination of the global and local adaptive binarization method at connected component level is proposed to binarize images. The performance of this method is extremely well but it is limited to binarize handwritten document images only.

Other non-threshold approaches have been reported and the results are promising. Howe [27] proposes a method based on the Laplacian energy of the image intensity. The energy function is minimized via a graph-cut computation. The method is efficient but parameter-dependent. In [28], Howe improves this method by tuning two key parameters adaptively and the performance is improved. Mishra et al. [29] define a energy function so that the quality of the binarization is inversely related to the energy value. They minimize this energy function to find the optimal binarization using an iterative graph cut scheme. In some literatures [30–33], the input images are divided into three classes: foreground, background, and uncertain. Then, they classify those uncertain pixels by applying the MRF model or other strategy on the other two categories of pixels. In a study published recently [34], some reasonable pre and post processes are used to deal with the broken and degraded document images. These methods which combine different types of image information and domain knowledge usually have high computational complexity.

3. Proposed method

This section details the proposed document image binarization method.

The flowchart is shown in Fig. 2. Given a document image, firstly, we compute the gradient map from the compensated image which is obtained through a background removal process. Then the SSP in the gradient map will be extracted from two different aspects: large magnitude selection using adaptive gradient binarization method and opposite orientation judgement based on the stroke width estimation. Finally, we use the multiple threshold vote based framework to decide whether each pixel belongs to text or not. The concrete procedures will be provided below.

[1] https://github.com/FuxiJia/DocumentBinarizationSSP.