ISauvola: Improved Sauvola's Algorithm for Document Image Binarization

Zineb Hadjadj[1,2(✉)], Abdelkrim Meziane[2], Yazid Cherfa[1], Mohamed Cheriet[3], and Insaf Setitra[2]

[1] University of Blida, Blida, Algeria
hadjadj_zineb@yahoo.fr, cherfa_yazid@yahoo.fr
[2] Research Center in Scientific and Technical Information (CERIST), Algiers, Algeria
{ameziane,isetitra}@mail.cerist.dz
[3] École de Technologie Supérieure, Montreal, QC, Canada
mohamed.cheriet@etsmtl.ca

Abstract. Binarization of historical documents is difficult and is still an open area of research. In this paper, a new binarization technique for document images is presented. The proposed technique is based on the most commonly used binarization method: Sauvola's, which performs relatively well on classical documents, however, three main defects remain: the window parameter of Sauvola's formula does not fit automatically to the image content, is not robust to low contrasts, and not invariant with respect to contrast inversion. Thus on documents such as magazines, the content may not be retrieved correctly. In this paper we use the image contrast that is defined by the local image minimum and maximum in combination with the computed Sauvola's binarization step to guarantee good quality binarization for both low and correctly contrasted objects inside a single document, without adjusting manually the user-defined parameters to the document content.

1 Introduction

1.1 Overview

One critical step of the analysis is to identify and retrieve foreground and background objects correctly; however it is not easy to binarize and find the best thresholds because of change of illumination or noise presumed issues.

For document images of a good quality, global thresholding [1–3] is capable of extracting the document text efficiently. But for document images suffering from different types of document degradation, adaptive thresholding, which estimates a local threshold for each document image pixel [4–8], is usually capable of producing much better binarization results.

Some binarization methods try to incorporate global or local approaches, like [9–12, 16]. Certain methods also have incorporated background estimation and normalization steps, like [11–13]. The image edges that can usually be detected around the text stroke boundary is also used in certain binarization methods, like [13–15].

© Springer International Publishing Switzerland 2016
A. Campilho and F. Karray (Eds.): ICIAR 2016, LNCS 9730, pp. 737–745, 2016.
DOI: 10.1007/978-3-319-41501-7_82

Because Sauvola's binarization is widely used in practice and gives good results on document images, this paper focuses on that particular method.

1.2 Sauvola's Algorithm and Issues

Sauvola's method [5] takes a grayscale image as input. Since most of document images are color images, converting color to grayscale images is required [17].

From the grayscale image, Sauvola proposed to compute a threshold at each pixel using:

$$T = m \times \left[1 + k \times \left(\frac{s}{R} - 1\right)\right] \tag{1}$$

where k is a user-defined parameter, m and s are respectively the mean and the local standard deviation computed in a window of size ω centered on the current pixel and R is the dynamic range of standard deviation ($R = 128$ with 8-bit gray level images). The size of the window used to compute m and s remains user-defined in the original paper.

The main advantages of Sauvola's method is that it performs relatively well on noisy and blurred documents [18] and its computational efficiency.

Due to the binarization formula (1), the user must provide two parameters $(\omega; k)$. Some techniques have been proposed to estimate them. [19] state that $\omega = 14$ and $k = 0.34$ is the best compromise for show-through removal and object retrieval quality in classical documents. [20] based the parameter research on Optical Character Recognition (OCR) result quality and found $\omega = 60$ and $k = 0.4$. [5, 18] used $\omega = 15$ and $k = 0.5$. Adjusting those free parameters usually requires an a priori knowledge on the set of documents to get the best results. Therefore there is no consensus in the research community regarding those parameter values.

In [21], a learning framework for the optimization of the binarization methods is introduced, which is designed to determine the optimal parameter values for a document image.

Sauvola's method suffers from different limitations among the following ones [22].

- *Missing low-contrast objects.*
- *Keeping textured text as is.*
- *Handling badly various object sizes.*
- *Spatial object interference.*

In the remainder of this paper, we present a method to overcome one of the four limitations of Sauvola's binarization mentioned previously, which is the *Missing of low-contrast objects.*

The rest of the paper is structured as follows. In Sect. 2 we first expose the general principle of the proposed method. In Sect. 3 we present some results of the proposed method applied to real documents and compare them to other methods' results. We conclude on the achievements of this work in Sect. 4.

2 Proposed Method

An improvement of Sauvola's algorithm is introduced in this section. This results in a three-step process: (1) An initialization map is extracted from the document image to identify high-probability text pixels; (2) Sauvola's algorithm is applied on the input document image; and, finally, (3) To produce the final binarization, we just have to detect in sauvola's binarization image the set of pixels overlapping with each text pixel of the initialization map.

In the next subsection, an initial binarization is estimated using the image contrast.

2.1 Step 1: Initialization Step

At this step, we use an initialization approach that is based on image contrast to identify high-probability text pixels. The used initialization first constructs a contrast image, evaluated by the local maximum and minimum, and then detects the high contrast image pixels which usually lie around the text stroke boundary.

- Contrast Image Construction

In the proposed technique, the used image contrast (Fig. 1(b)) is calculated based on the local image maximum and minimum [15] as follows:

$$D(x,y) = \frac{f_{max}(x,y) - f_{min}(x,y)}{f_{max}(x,y) + f_{min}(x,y) + \epsilon}, \tag{2}$$

where $f_{max}(x,y)$ and $f_{min}(x,y)$ refer to the maximum and the minimum image intensities within a local neighborhood window. In the implemented system, the local neighborhood window is a 3×3 square window. The term ϵ is a positive and very small number, which is added in case the local maximum is equal to 0.

The image contrast in (2) minimizes the image background and brightness variation properly. In particular, the numerator captures the local image difference that is similar to the traditional image gradient. The denominator (the normalization term) is used to avoid an artifact of uneven background and lowers the effect of the image contrast and brightness variation [15].

- High Contrast Pixels Detection

The purpose of the contrast image construction is to detect the desired high contrast image pixels lying around the text stroke boundary. As described in the last subsection, the constructed contrast image has a clear bimodal pattern where the image contrast around the text stroke boundary varies within a small range but is obviously much larger compared with the image contrast within the document background. We therefore detect the desired high contrast image pixels by using Otsu's global thresholding method (Fig. 1(c)).

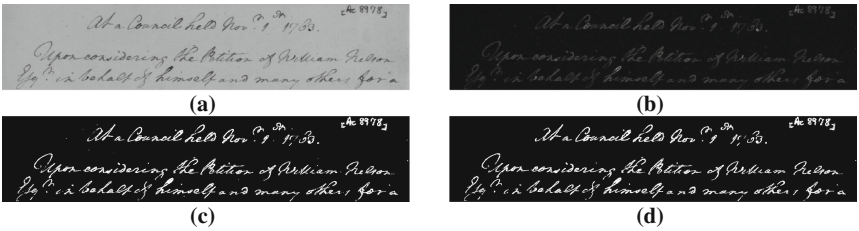

Fig. 1. High contrast pixel detection: (a) input image, (b) image contrast, (c) high contrast image pixels and (d) is (c) after morphological opening operation

To remove the small objects, we use the morphological open binary image operator[1] (Fig. 1(d)). As can be seen in Fig. 1(d) some faint characters or low contrasted text pixels was suppressed but this issue will be solved at the third step.

2.2 Step 2: Sauvola's Binarization Step

At this step, Sauvola's thresholding described in Sect. 1.2 is performed on the input document image.

In our experiments, we found that the value of R in Eq. (1) has a very small effect on the binarization quality while the values of k and window size affect it significantly. Low contrasted objects may be considered as textured background or show-through artifacts due to the threshold formula and may be removed or partially retrieved. A low value of k parameter can help retrieving low-contrasted objects but since it is set for the whole document, it also alters other parts of the result: correctly contrasted objects are thicker in that case, possibly causing unintended connections between components. This is due to the fact that background noise and artifacts are usually poorly contrasted and are retrieved as objects.

The size of the window is an important parameter to get good results, too low a value may lead to broken characters and/or characters with holes whereas too large a value may lead to bold characters. Its size must depend on the contents of the document.

An optimal combination of k and ω will produce a good binary image. In our experiments we choose a low value for k parameter to detect all the text pixels (low and correctly contrasted) and a low value for ω parameter to reduce the overlapping between characters.

2.3 Step 3: Sequential Combination

At this step, we use a sequential combination between the contrast image and Sauvola's binarization image to obtain the final binary result.

[1] Numerically, this is done by using the function **bwareaopen** of Matlab.

The sequential combination consist in detecting in Sauvola binarization image the set of pixels overlapping with each text pixel of the high contrast image as described in Algorithm 1.

Algorithm 1. Step 3: Sequential combination between the high contrast image and Sauvola's binarization image.

Require: The high contrast image I_C (constructed at step 1) and Sauvola binarization image I_S (constructed at step 2)
Ensure: The final binary result I_B
1: **for** all pixel p in I_C: **do**
2: **if** I_C(p)=true **then** //p is part of an object in I_C
3: Detect the set of pixels overlapping with p in I_S.
4: **end if**
5: **end for**
6: Store the new binary result to I_B.

The proposed method described was implemented in Matlab; the results are presented and discussed in the following section.

3 Experiments and Discussion

The described method has been tested on the document images used in the Document Image Binarization Contests (DIBCO) that suffer from different types of document degradation. We also compare our method with other well-known binarization methods including Sauvola's thresholding method [5].

Multiple tests performed on document images have demonstrated that the following parameters: $\omega = 15 \times 15$, $R = 128$ as recommended in [5] and $k = 0.01$ give the best binarization results. A low value of k parameter can help retrieving low-contrasted objects, since it is set for the whole document, it also alters other parts of the result: lot of background noise and artifacts are retrieved as objects but our proposed sequential combination (the third step) can suppress the noise efficiently because it suppresses the contrast of the document background through the normalization as described in Sect. 2.1

The binarization results in Figures below show the superior performance of the proposed thresholding technique.

Figure 2(a) illustrates a faint characters degraded document image. Our binarization technique first constructs a contrast image by the local image maximum and minimum and then extracts the high contrast image (Fig. 2(b)) which is used to suppress the background noise and artifacts. Then Sauvola's algorithm is applied on the input image (Fig. 2(c)) to detect all the text pixels (low and correctly contrasted). After that, we combine sequentially between the two results by detecting in Sauvola binarization image the set of pixels overlapping with each text pixel of the high contrast image to produce the final binarization. As can be seen in Fig. 2(f) (the final result) the faint characters are reasonably well detected by using our method. On the other hand,

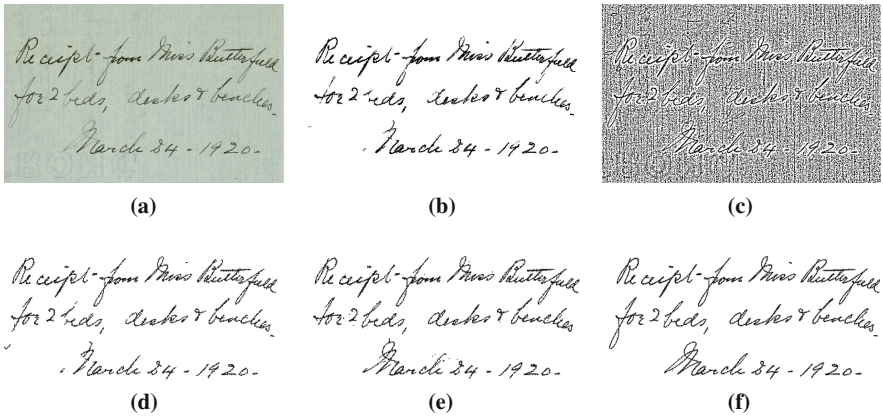

Fig. 2. Image HW2 from the DIBCO'11 test dataset: (a) input image, (b) high contrast image, (c) binarization result obtained using Sauvola's method, (d) Su's method (ranked 2nd in DIBCO'11), (e) Howe's method (ranked 3rd in DIBCO'11) and (f) the result of the proposed method

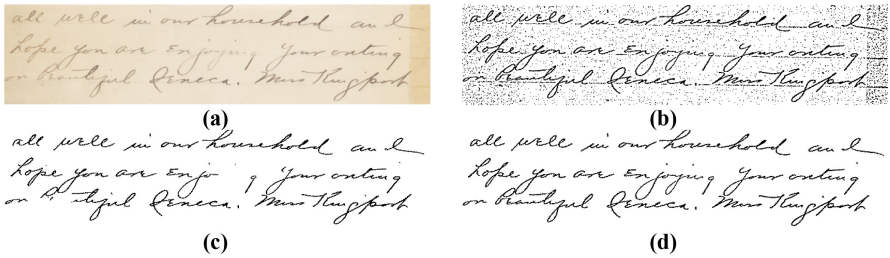

Fig. 3. Image H12 from the HDIBCO'12 test dataset: (a) input image, (b) binarization results obtained using Sauvola's method, (c) Lelore's method (ranked 2nd in H-DIBCO'12) and (d) the proposed method

Sauvola's method produces a lot of noise due to the variation within the document background, Su's method and Howe's method cannot detect some weak characters .

Figure 4(a) illustrates a bleed-through degraded document image, as can be seen in Fig. 4(c) the bleed-through is reasonably well removed by using our method. The proposed method can suppress more noise than Sauvola's method because it suppresses the contrast of the document background through the normalization in the initialization step. As a comparison, Sauvola's method simply classifies dark background pixels as the text pixels improperly.

Figures 3(a) and 5(a) illustrate a faint characters degraded document image. Figure 5 shows that the proposed technique is tolerant to the variations in document contrast and able to binarize faint characters and badly illuminated image with little background noise where some other methods may either introduce a certain amount of noise or fail to detect the document text with a low image contrast shown in Fig. 3(c).

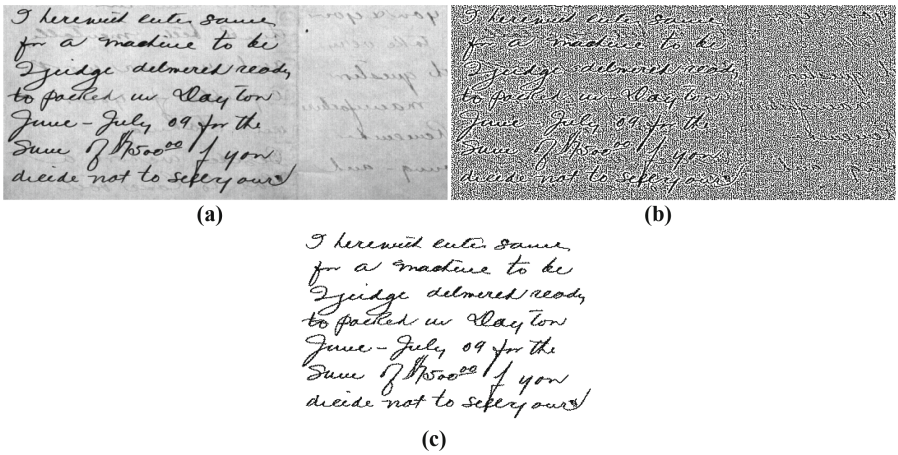

Fig. 4. Image HW4 from the DIBCO'13 test dataset: (a) input image, (b) binarization results obtained using Sauvola's method and (c) the proposed method

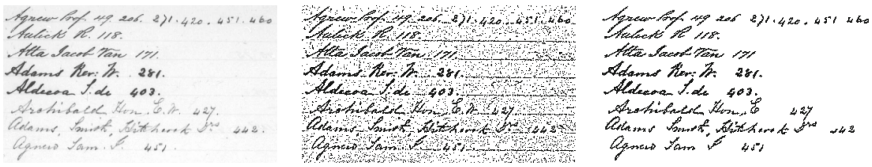

Fig. 5. Image H04 from the HDIBCO'10 test dataset: (a) input image, (b) Sauvola's method and (c) the proposed method

Figures 2, 3, 4 and 5 further show four document binarization examples. As shown, our proposed method extracts the text properly from document images that suffer from different types of degradation. On the other hand, Sauvola's method often produces a certain amount of noise due to the variation within the background.

4 Conclusion and Future Prospects

This paper presents an efficient historical document image binarization technique that is efficient against different types of document degradation such as faint characters and uneven illumination. The proposed technique makes use of Sauvola's algorithm and the image contrast that is evaluated based on the local minimum and maximum. Such a combined method leads, as shown in the experiments, to high accuracy when applied to historical document images, with a variety of degradations. The proposed method succeeds indeed to capture low and correctly contrasted objects inside a single document. However, the performance of our binarization method is limited in case of both small and large objects in a same document; Sauvola's method fails to retrieve all

objects correctly because its window parameter is defined for an image as a whole. In future work we will focus on handling various object sizes.

References

1. Otsu, N.: A thresholding selection method from gray-level histogram. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
2. Kapur, J.N., Sahoo, P.K., Wong, A.K.C.: A new method for gray-level picture thresholding using the entropy of the histogram. Graph. Image Process. **29**, 273–285 (1985)
3. Kittler, J., Illingworth, J.: Minimum error thresholding. Pattern Recognit. **19**(1), 41–47 (1986)
4. Niblack, W.: An Introduction to Digital Image Processing. Prentice Hall, Englewood Cliffs (1986)
5. Sauvola, J., Pietikainen, M.: Adaptive document image binarization. Pattern Recognit. **33**(2), 225–236 (2000)
6. Bernsen, J.: Dynamic thresholding of grey-level images. In: Proceedings of the Eighth International Conference on Pattern Recognition, Paris, France, pp. 1251–1255, October 1986
7. Wolf, C., Jolion, J.M.: Extraction and recognition of artificial text in multimedia documents. Pattern Anal. Appl. **6**(4), 309–326 (2003)
8. Feng, M.L., Tan, Y.P.: Contrast adaptive binarization of low quality document images. IEICE Electron. Express **1**(16), 501–506 (2004)
9. Kim, I.K., Jung, D.W., Park, R.H.: Document image binarization based on topographic analysis using a water flow model. Pattern Recogn. **35**(1), 265–277 (2002)
10. Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. Pattern Recogn. **39**(3), 317–327 (2006)
11. Lu, S., Su, B., Tan, C.L.: Document image binarization using background estimation and stroke edges. Int. J. Doc. Anal. Recogn. **13**(4), 303–314 (2010)
12. Ntirogiannis, K., Gatos, B., Pratikakis, I.: A combined approach for the binarization of handwritten document images. Pattern Recogn. Lett. - Spec. Issue Front. Handwrit. Process. **35**, 3–15 (2012). doi:10.1016/j.patrec.2012.09.026
13. Moghaddam, R.F., Cheriet, M.: RSLDI: restoration of singlesided low-quality document images. Pattern Recogn. **42**(12), 3355–3364 (2009)
14. Howe, N.: Document binarization with automatic parameter tuning. Int. J. Doc. Anal. Recogn. **16**, 247–258 (2012)
15. Su, B., Lu, S., Tan, C.L.: Binarization of historical handwritten document images using local maximum and minimum filter. In: International Workshop on Document Analysis Systems, pp. 159–165, June 2010
16. Hadjadj, Z., Meziane, A., Cheriet, M., Cherfa, Y.: An active contour based method for image binarization: application to degraded historical document images. In: ICFHR 2014, Crete, Greece, pp. 655–660 (2014). doi:10.1109/ICFHR.2014.115
17. Moghaddam, R.F., Cheriet, M.: A multi-scale framework for adaptive binarization of degraded document images. Pattern Recogn. **43**(6), 2186–2198 (2010)
18. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. J. Electron. Imaging **13**, 146–165 (2004)
19. Badekas, E., Papamarkos, N.: Automatic evaluation of document binarization results. In: Sanfeliu, A., Cortés, M.L. (eds.) CIARP 2005. LNCS, vol. 3773, pp. 1005–1014. Springer, Heidelberg (2005)

20. Rangoni, Y., Shafait, F., Breuel, T.M.: OCR based thresholding. In: Proceedings of IAPR Conference on Machine Vision Applications, pp. 98–101 (2009)
21. Cheriet, M., Moghaddam, R.F., Hedjam, R.: A learning framework for the optimization and automation of document binarization methods. Comput. Vis. Image Underst. (CVIU) **117** (3), 269–280 (2013)
22. Lazzara, G., Géraud, T.: Efficient multiscale Sauvola's binarization. Int. J. Doc. Anal. Recogn. **17**(2), 105–123 (2014)