Chinese Historic Image Threshold Using Adaptive K-means Cluster and Bradley's

Zhi-Kai Huang(✉), Yong-Li Ma, Li Lu, Fan-Xing Rao, and Ling-Ying Hou

College of Mechanical and Electrical Engineering, Nanchang Institute of Technology, Nanchang 330099, Jiangxi, China
huangzhik2001@163.com

Abstract. Resorting to extraction text techniques for Chinese heritage documents becomes an increasing need. Historic documents such as Chinese calligraphy usually were handwritten or scanned in low contrast so that an automatic optical character recognition procedure for document images analysis is difficult to apply. In this paper, we present a historic document image threshold based on a combination of Bradley's algorithm and K-means. An adaptive K-means cluster as a pre-processing methods for document image has been used for automatically grouping the pixels of a document image into different homogeneous regions. In Bradley's methods, every image's pixel is set to black if its brightness is T percent lower than the average brightness of surrounding pixels in the window of the specified size, otherwise it is set to white. Finally, text bounding boxes are generated by concatenating neighboring word clusters with mathematical morphology method. Experimental results show that this algorithm is robust in dealing with non-uniform illuminated, low contrast historic document images in terms of both accuracy and efficiency.

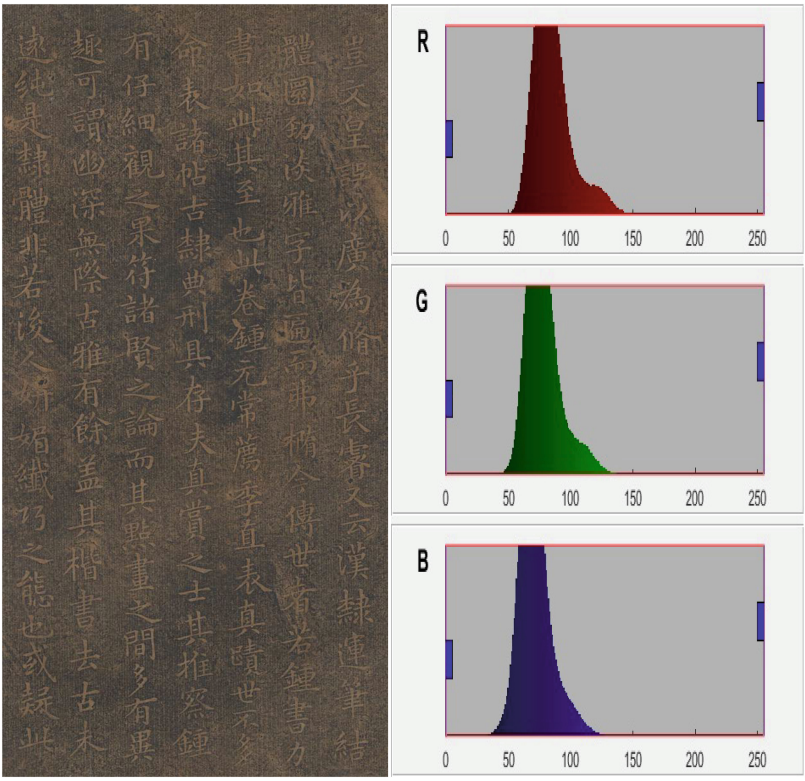
Keywords: Chinese historical image · K-means cluster · Bradley's method

1 Introduction

The digital rubbing is a novel approach to promote and pass on Chinese traditional arts, as well as a new idea to protect stone relics. Historical printed documents, such as old books and rubbings, are being digitized and made available through software interfaces such as web-based libraries, for instance, there is a large collection of Chinese rubbing database keeps in UC Berkeley east Asian library [14, 15]. These documents are challenging for OCR (Optical Character Recognition, OCR) because it use non-standard fonts and suffer from printing noise, artifacts due to aging, varying kerning (space between letters), varying leading (space between lines), frequent line break hyphenation, and other image problems due to the conversion from print-to-microfiche-to-digital [1]. Segmenting heritage documents images into text from background is a crucial pre-processing step for automated reading of historical documents. A Chinese historical rubbing image has been shown in Fig. 1. Many stone texture patches have shown in background of rubbing image because of the nature factor (all characters have been carved out of stone, ink rubbings has reproduced from

D.-S. Huang et al. (Eds.): ICIC 2016, Part III, LNAI 9773, pp. 171–179, 2016.
DOI: 10.1007/978-3-319-42297-8_17

that stone). The histogram has showed in three bands with a unimodal distribution, it is difficult to calculate the specific threshold.



(a) (b)

Fig. 1. Historical handwritten document image with low contrast and histogram (a) Original image, (b) The Histogram of original image

Document image binarization is a key step from the image processing to decreases the computational load and enables the utilization of the simplified analysis methods compared to 256 levels of grey-scale or color image information, which is also a basic technique in the computer vision. It makes the high-level computer vision tasks possible, so it plays an important role in the technology of the document image processing [2]. A number of promising techniques for document image binarization were implemented at the different literatures [3]. Generally, the methods that deal with document image binarization may be broadly categorized either globally or locally. Whether it is global or local binarization, threshold choice is a sensitive function of the local reflectance map. Specially, for low contrast scanned historic document image, it is difficult to improve on a fixed threshold centered between the extreme observed values, since too low a value will swamp the difference map with spurious changes, while too

high a value will suppress significant change [11].The most classical threshold algorithm that is Otsu's method [5], which maximizes the values of class variances to get optimal threshold. Maximum entropy was firstly proposed by Pun [12]. The purpose is to divide the gray-level histogram of image into separate classes and make maximum total entropy of all kinds of classes. An automatic threshold algorithm based on an iterative threshold selection is employed in the study [6–8]. At iteration [10], a new threshold Tn is established using the average of the foreground and background class means. The iterations terminate when the changes $|Tn–Tn + 1|$ become sufficiently small [9]. Sawaki and Hagita demonstrated another specialized binarization method for textured and reverse-video (white-on-black) Japanese headlines [13]. Their method is based on the complementary relationship between characters and their backgrounds as indicated by similarity measure for black and white runs. From Fig. 1(b), the gray-level distribution did not revealed bimodal, it is important to take into consideration the amplitude transfer function of the specific scanner, as well as the spatial and gray-scale characteristics of the image. That is to say, the pixels on an image are highly correlated, i.e. the pixels in the immediate neighborhood possess nearly the same feature data. Therefore, the spatial relationship of neighboring pixels is an important characteristic that can be of great aid in imaging segmentation. Cluster techniques have taken advantage of this spatial information for image segmentation [4].

In this paper, an adaptive foreground and background clustering (FBC) approach to document image binarization, each pixel is assigned to a foreground cluster or a background cluster. The algorithm is based on adaptive K-means algorithm, where the cluster means are updated each time a data point is assigned to a cluster. Since only one background and one foreground is assumed ($K = 2$), that is to say, only two clusters are considered, which makes the overall implementation easier. Following by that, a median filter has been employed for salt and pepper noise removing. Finally, a Connected components labeling technique is devised to locate possible positions of Chinese character.

2 Proposed Work

Figure 1 shows a scan of a page from a Chinese historical book that the rubbing image gets darker shows a low-contrast image with its histogram. As it can be observed from the luminance histogram, all the values gather in the left of the three bands, so it is impossible to reliably locate a local minimum between histogram valleys.

K-means clustering is one of the popular algorithms in clustering and segmentation. It treats each image pixel (with R,G,B values) as a feature point having a location in space. The basic K-means algorithm then arbitrarily locates, that number of cluster centers in multidimensional measurement space. Each point is then assigned to the cluster whose arbitrary mean vector is closest. The procedure continues until there is no significant change in the location of class mean vectors between successive iterations of the algorithms. Firstly, we use an adaptive K-means cluster based segmentation to improve the performance of threshold image.

2.1 Bradley's Algorithm

The main idea in Bradley's algorithm is that compute the sum of real numbers $f(x, y)$ (for instance, pixel intensity) over a rectangular region of the image. It could be called as integral images. To compute the integral image, we store at each location, $I(x, y)$, the sum of all $f(x, y)$ terms to the left and above the pixel (x, y). This is accomplished in linear time using the following equation for each pixel (taking into account the border cases),

$$I(x, y) = f(x, y) + I(x - 1, y) + I(x, y - 1) - I(x - 1, y - 1) \quad (1)$$

After that, compute the $s \times s$ average using the integral image for each pixel in constant time and then perform the comparison. If the value of the current pixel is t percent less than this average then it is set to black, otherwise it is set to white. The pseudo code for Bradley's algorithm has been showing as following:

Input image In, output binary image out, image width w and image height h.

```
1: for i = 0 to w do
2:  sum ← 0
3:  for j = 0 to h do
4:    sum ← sum+in[i, j]
5:    if i = 0 then
6:     intImg[i, j] ← sum
7:    else
8:    intImg[i, j] ← intImg[i−1, j]+sum
9:    end if
10:  end for
11: end for
12: for i = 0 to w do
13:   for j = 0 to h do
14:    x1 ← i−s/2 {border checking is not shown}
15:    x2 ← i+s/2
16:    y1 ← j−s/2
17:    y2 ← j+s/2
18:    count ← (x2−x1)×(y2−y1)
19:                                    sum                                    ←
     intImg[x2,y2]−intImg[x2,y1−1]−intImg[x1−1,y2]+intImg[x1−1,y1−1]
20:     if (in[i, j]×count) ≤ (sum×(100−t)/100) then
21:     out[i, j] ← 0
22:    else
23:   out[i, j] ← 255
24:  end if
25: end for
26: end for
```

2.2 Overview of the Binarization Technique

Because there are some amount of 'salt & pepper' noise exist in the document image after Bradley's binarization, median filtering is conducted. Finally, a morphology-based technique is devised to locate possible positions of Chinese Character. The detail process is shown in Fig. 2.

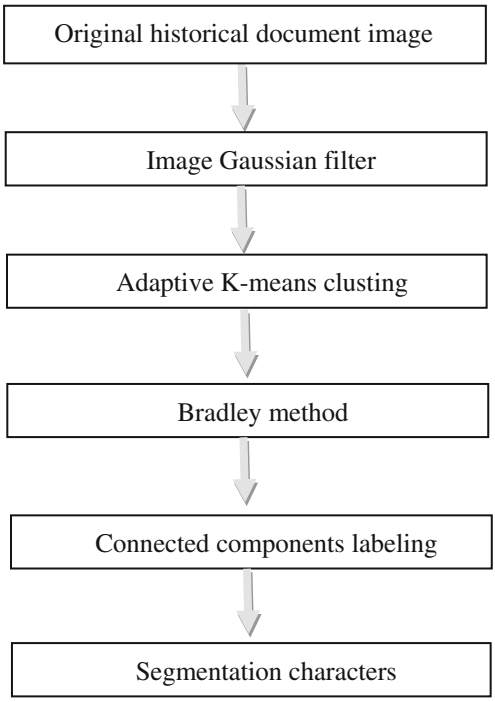
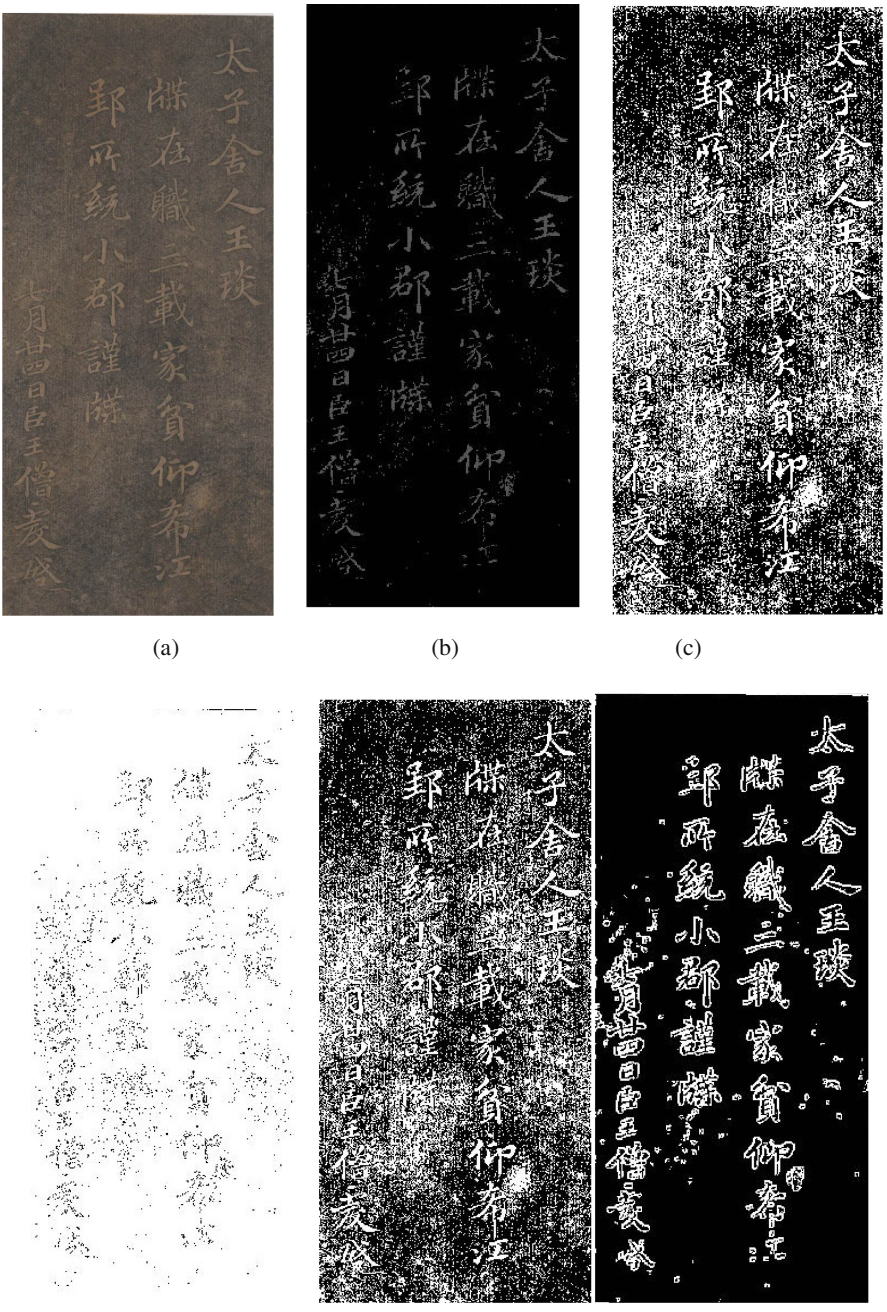


Fig. 2. Flow diagram of document binarization system

3 Experiments Performed on Chinese Rubbing Images

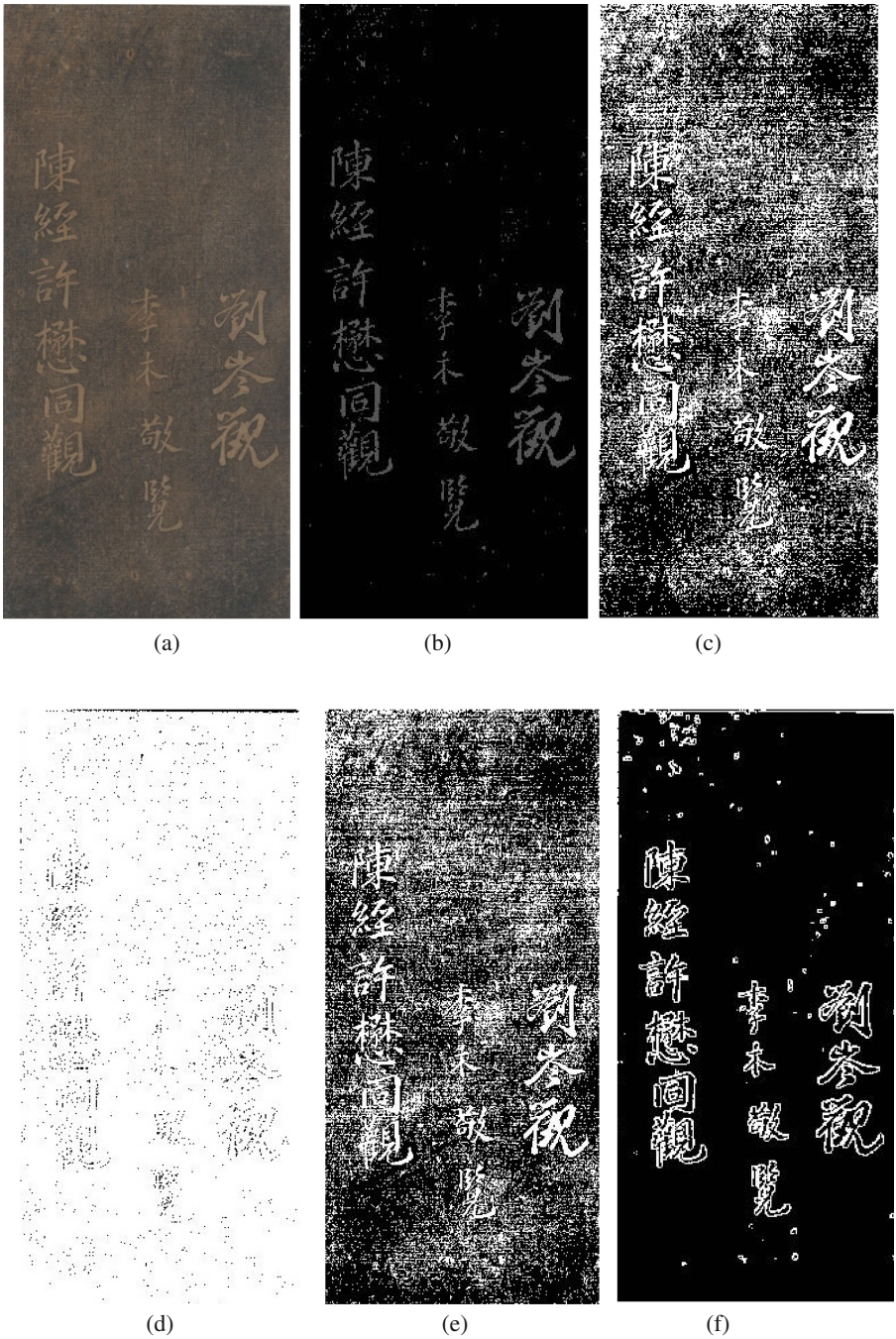
The algorithm is implemented in MATLAB. The algorithm is tested with many scanned Chinese rubbing document images, which contained characters of different fonts and size. The parameters in the Bradley's binarization experiments are set as follows: W = 15, H = 15, and T = 5. The median filtering using MATLAB's medfilt2 function, we have used neighborhoods of size 3-by-3 to remove noising. There chosen two examples of our adaptive threshold result are presented in Figs. 3 and 4. In order to compare our method with other different algorithm, the results of Sauvola's, Ostu's and Isodata's method are shown in same figure, also. Figures 3 and 4 illustrate a text example with a very low contrast. Our method is able to segment most of the text in the image.



(a) (b) (c)

(d) (e) (f)

Fig. 3. (a) Original image; (b) Adaptive K-means cluster image; (c) Isodata method; (d) Sauvola's method; (e) Ostu's method; (f) Our method



(a) (b) (c)

(d) (e) (f)

Fig. 4. (a) Original image; (b) Adaptive K-means cluster image; (c) Isodata method; (d) Sauvola's method; (e) Ostu's method; (f) Our method

We tried to look for existing methods to deal with Chinese historic image binarization problem. After many experiments have been implemented, our technique can provide near perfect segmentation despite the very low illumination in the image. Sauvola's technique fails at the characters detection. Ostu's global technique and isodata method keep more noising in image.

4 Conclusion

In this paper, a Chinese rubbing document image binarization algorithm is developed from low contrast Chinese rubbing images. The proposed scheme is developed based on an adaptive K-means combined Bradley operation. We tested our scheme on many different Chinese rubbing, and obtained encouraging results. Because the validation procedure should handle the variety of handwritten characters and the ambiguity in the distinction of characters, extracted them from images is not always easy and is still a topic of future research.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grant No. 61472173), the grants from the Educational Commission of Jiangxi province of China, No. GJJ151134.

References

1. Gupta, M.R., Jacobson, N.P., Garcia, E.K.: OCR binarization and image pre-processing for searching historical documents. Pattern Recogn. **40**(2), 389–397 (2007)
2. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recogn. **33**(2), 225–236 (2000)
3. Yan, H.: Unified formulation of a class of image thresholding techniques. Pattern Recogn. **29**(12), 2025–2032 (1996)
4. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
5. Otsu, N.: A threshold selection using gray level histograms. IEEE Trans. Syst. Man Cybernet. **9**, 62–69 (1979)
6. Bradley, D., Roth, G.: Adaptive thresholding using the integral image. J. Graph. GPU Game Tools **12**(2), 13–21 (2007)
7. Wellner, P.D.: Adaptive thresholding for the DigitalDesk. Xerox, EPC1993-110 (1993)
8. Pappas, T.N.: An adaptive clustering algorithm for image segmentation. IEEE Trans. Signal Process. **40**(4), 901–914 (1992)
9. Chang, C.I., Du, Y., Wang, J., et al.: Survey and comparative analysis of entropy and relative entropy thresholding techniques. In: IEE Proceedings - Vision, Image and Signal Processing, IET, vol. 153(6), pp. 837–850 (2006)
10. Sezgin, M.: Survey over image thresholding techniques and quantitative performance evaluation. J. Electron. Imaging **13**(1), 146–168 (2004)
11. Huang, Z.K., Chau, K.W.: A new image thresholding method based on Gaussian mixture model. Appl. Math. Comput. **205**(2), 899–907 (2008)

12. Hayes, B., Wilson, C.: A maximum entropy model of phonotactics and phonotactic learning. Linguist. Inq. **39**(3), 379–440 (2008)
13. Mori, M., Sawaki, M., Yamato, J.: Robust character recognition using adaptive feature extraction. In: 23rd International Conference on Image and Vision Computing New Zealand, IVCNZ 2008, pp. 1–6. IEEE (2008)
14. http://www.lib.berkeley.edu/EAL/stone/rubbings.html
15. http://vc.lib.harvard.edu/vc/deliver/home?_collection=rubbings