Improved Line Segmentation Framework for Sundanese Old Manuscripts

E Paulus, M Suryani, S Hadi
Department of Computer Science, University of Padjadjaran, Jalan Raya Bandung Sumedang KM21 Jatinangor 45363, INDONESIA
E-mail: erick.paulus@unpad.ac.id, mira.suryani@unpad.ac.id, setiawanahdi@unpad.ac.id

Abstract. Line segmentation can be a useful process for further text segmentation. There are some certain line segmentation framework that use binarization method as an initial step. But binarization process is still facing a major challenge, especially on old document palm-leaf manuscripts. As the quality of the image has varying degrees of noises in the non-text region. Seam Carving method, one of line segmentation methods that uses binarization-free approach, can be an alternative solution. However, this method can separate the incorrect text line on small element text located at the bottom or at the top of a main character contour. Therefore, an improvement on line segmentation framework is proposed by using hybrid binarization and its implemented on the smallest energy function to separate out the text-lines. The proposed framework have been evaluated on 44 Sundanese old manuscript images that consist of true color and binary images. The evaluation matrix shows that this framework can improve Niblack binarization process up to 50%. In addition, our framework does not only generate the number of text-lines to come near to the number of target lines, but it also can separate the text-lines well on small element text. Overall, the expected result can in the end be produced from the proposed line segmentation framework.

1. Introduction

Preservation programs on old manuscripts have become important issues for the government, library, and society. Digitization process, which is one of preservation types, have produced many manuscript images. OCR system is needed to extract the text contained in the manuscript images. To recognize the text pattern, binarization and segmentation processes are required as fundamental steps. Some line segmentation methods can produce text line using binarization-free approach. It means that we can separate text line directly from true color image. However, whether or not line segmentation using binarization-free approach can deliver the best result compared to line segmentation using binarization approach is still uncertain. Therefore, we evaluate the performance of Seam Carving method. Then, we construct binarization and line segmentation framework to obtain better solution in separating text line. In manuscript image analysis, binarization process has an important role to transform true color or grey scale image into binary image[1]. There are some challenges to binarize the old document image such as low contrast, stain, smear, shadow, and irregular illumination[1-2]. Diversiform binarization algorithm has been formulated to convert true color image into binary images using global, local, or hybrid thresholding [3-9]. Five steps of binarization process are proposed by Ntogas.
[6] for Byzantine old manuscripts. For old manuscripts, many studies show that binarization method using hybrid thresholding is better [11]. There is binarization algorithm that uses whale optimization approach and implement fuzzy c-means as objective function[8], contour model[3] and stroke connectivity[9]. Local thresholding such as Niblack, Sauvola[5] and Howe[15] can also give contribution in binarization. The binarization performance of Sauvola and Howe show better result than Niblack. But Niblack method can keep the solid contour of the text, though a lot of subtle noise still remain on binary image. Hence, we improve hybrid binarization framework using the advantage of Niblack method, edge detection, color map, and filtering method. After binarization process, we can continue to line segmentation process[12]. This step eases segmentation process for the character, syllable and word. Projection profile method[13], as the simplest line segmentation, can identify the line of a text by using histogram profile. The profile presents valley points at line boundaries and the location of these minimum points mark the line boundaries. This method runs well if the input image is binary image and there is no waved text. Adaptive projection profile [14] is used to deal with waved text, by dividing the image into some columns. Seam Carving method implements the minimum energy function to separate the lines of a text. The advantage of Seam Carving method works well on the true color image and waved text. But, this method does not work well on some small element text located at the bottom or at the top of the main character. Figure 1 shows the incorrect line segmentation. In this paper, binarization and line segmentation frameworks are exhibited in detail. Systematically, the paper is divided into sections. The next section of this paper (the second section) describes the collection of Sundanese old manuscripts and the challenges for binarization and line segmentation. Following this, our proposed binarization and line segmentation frameworks are shown on the third section. Some experiments and results of our proposed framework are explained in the fourth section. The last section describes the conclusions and some potential research for future study.

Figure 1. The examples of incorrect line segmented both on color and binary image.

2. The Condition and Challenges of Sundanese Old Manuscripts
Several mediums are used in old manuscripts such as paper, palm leaf, metal and stone. Generally, Sundanese old manuscripts in museums are of many mediums. But, there are some other places that keep both the original and duplication of the manuscripts in a private place. Kabuyutan Ciburuy, one of Indonesia’s private heritage place, stores the collection of Sundanese old manuscripts. These collections, known as Sundanese Lontar uses palm leaf as its writing medium. The length and width of Sundanese Lontar is about 25 to 45 cm and 10 to 15 cm respectively. Each page consists of four to six lines, and it is estimated there are 15 to 20 words in each line. The Sundanese Lontar has a hole in the center and it is rope-bound. It consists of a variety of writings such as Ramayana epoch, farm and medical formulae, and social life. This made the Sundanese Lontar extremely valuable for Sundanese people. The quality of Sundanese old manuscript images are poor, as they contain various noises such as smears, non-uniform illumination, shadow, and many random noises. All of the characteristics are the challenges for every step in image processing including binarization and text line segmentation.
3. The Proposed Line Segmentation Framework

Our proposed framework, is of two part steps, namely: pre-processing and line segmentation process. We also proposed novel filtering to increase the performance of Niblack binarization. This framework is designed for the designated Sundanese old manuscripts.

3.1. Pre-processing

The goal of pre-processing step is to minimize the noise that is distributed on the original image and transform it to binary image. The first step, We convert RGB color map into the HSV color map and Grayscale. The second step is to remove Non-Lontar Area. In digitizing process, the photographer used fabric (dark or light color) as background so we needed to remove the non-lontar area. We used hue or saturation channel to transform it into white color.

The third step, we evaluated several thresholding methods such as Otsu[6], Niblack[4][10], Sauvola[5], and Howe[15] to our framework. Then, the result showed that Niblack is suitable for our framework. Even though Niblack still leaves noise everywhere. But, the contour of the text is still solid. There were no missing texts. This advantage of Niblack was chosen and implemented in our framework. The binarization process can be formulated on equation (1).

\[ T = m + k \times s \]  

where \( s \) stands for standard deviation, and \( m \) for average of color intensity \( I \) in the sub window. Also, \( k \) is a constant a value in \([-1,0) \) and \((0,1]\). Based on the experiment, \( k = + 0.2 \) is appropriate for white object detection and \( k = -0.2 \) is appropriate for black object detection.

The last step, Filtering was used to remove noise that is distributed in the digital image area. We proposed a novel filtering to improve Niblack method, as presented in equation (2) and (3). The goal of this filtering is to remove the tiny group of noise. It is represented by the sum of black area which is less than a half of white area.

\[ E(i, j) = \sum_{x=-r}^{r} \sum_{y=-r}^{r} f(i+x, j+y) \]  

\[ bw_{filter}(i, j) = \begin{cases} 1 \text{ (white color)} & , \text{if } E(i, j) \geq \left( 2 \times r + 1 \right)^2 / 2 \\ 0 \text{ (black color)} & , \text{otherwise} \end{cases} \]  

where \( r \) is a half-length \( (N) \) of sub window \( (N \times N) \); \( E \) is the cumulative sum of white region.

3.2. Line Segmentation

The basic method for line segmentation is projection profile[13]. Projection profile can be processed vertically or horizontally. Vertical projection profiles can segment word or syllable while horizontal projection profile can segment text line. Nowadays, there are several line segmentation algorithm adopted from Seam Carving method [16-18]. Some studies explain that this algorithm is free from binarization. Hence, it can be an alternative solution to our old manuscript segmentation. The following is a segmentation algorithm.

The first process, the image was divided vertically into several pieces. Then, we implemented Sobel Operator as edge detection. After that, we calculate the smoothed horizontal projection profile for each peace. Finally, we found and connected the nearest local maximum of each piece.

The second process is to compute the Separator of Text-line. This process implements modified Seam Carving procedure. We calculated Energy Function as initial step. Then, we calculated the cumulative of minimum energy \( M \). After that, we calculated the optimal path from cumulative energy \( M \). The high-energy regions represent the foreground (text component) and the low-energy regions represent the background (non-text component). The sample image of separating text-line is shown in Figure 1. Sometimes there are some text components which were not included in their lines.
4. Experiments and Result

We used datasets which consist of 22 true color images of Sundanese old manuscripts and 22 ground truth binary images. True color images consisted of 12 true color images from Kropak 18 and 10 true color images from Kropak 22. In general, every image contains four lines. Then, we used PixLabelers[19] tools to construct the ground truth (GT) images. After that, we segmented it to become text-line GT images using Alethea Lite[20] tools and built an application program of image cropping based on XML file using Scilab[21]. This experiment implemented the evaluation matrix that was used in ICDAR contest in 2013.

4.1. Experiment I - Pre-processing Phase

The first sub framework is proposed to increase the performance of binarization for Sundanese manuscript. We investigated some common binarization methods such as Otsu, Niblack, Sauvola, and Howe. Figure 2 presents the result of pre-processing phase. We found that Niblack binarization method resulted in many noises. Apart from that, the contour text is still visually solid. Therefore, we added some filtering method to improve the pre-processing phase. Our proposed framework succeeded in getting the highest score on binarization evaluation, as presented in table 1–2.

![Figure 2. Example of binarized images (up to bottom) using the methods of Otsu[11], Niblack[4], Sauvola[5], Howe[15], and proposed scheme.](image)

Table 1. The binarization evaluation matrix based on ICDAR competition[11] for 3309.tif

| Methods       | Precision | Recall | F-m (%) | pF-m (%) | NRM | PSNR | DRD | MPM (x1000) |
|---------------|-----------|--------|---------|----------|-----|------|-----|-------------|
| Otsu          | 0.71      | 0.19   | 30.37   | 31.76    | 0.43| 4.92 | 25.38| 109.97      |
| niblack       | 0.24      | 0.71   | 35.74   | 36.43    | 0.27| 5.98 | 43.76| 151.99      |
| sauvola       | 0.36      | 0.46   | 40.74   | 43.20    | 0.31| 8.80 | 20.60| 64.08       |
| Howe          | 0.30      | 0.56   | 39.38   | 40.49    | 0.29| 7.71 | 27.58| 62.34       |
| OurProposed   | 0.46      | 0.71   | 56.03   | 57.53    | 0.19| 9.61 | 16.33| 18.01       |

Table 2 describes the binarization performance for Kropak 18 and Kropak 22 using some common binarization methods. For Kropak 22, Niblack, Sauvola, Howe and our proposed method result in the average F-measure above 50%. But, the highest value of F-measure was obtained by our proposed method, which amounted to 58,51, while Sauvola acquired the highest PSNR of about 11.79, followed by our proposed method of about 10.86. In this case, the Sauvola binarized image is closer to ground truth image. The F-measure average of Kropak 18 is still below 50% because of the poor physical condition of the manuscripts: smudges, black spot, and low contrast. Furthermore, the MPM average of Kropak 18 decreased from 0.146 point to 0.018 point. It shows that the resulted image is closer to ground truth image, but the result is not
Table 2. The binarization evaluation matrix based on ICDAR Competition

| Methods      | Koropak 18 |         | Koropak 22 |         |
|--------------|------------|---------|------------|---------|
|              | F-m (%)    | pF-m (%)| PSNR       | MPM x1000 | F-m (%)  | pF-m (%)| PSNR       | MPM x1000 |
| Otsu         | 21.44      | 23.78   | 4.23       | 158.68   | 26.03    | 26.21   | 4.11       | 304.41    |
| Niblack      | 29.12      | 29.56   | 5.78       | 146.18   | 50.54    | 51.32   | 8.72       | 68.60     |
| Sauvola      | 30.82      | 32.30   | 8.88       | 64.46    | 51.65    | 55.62   | 11.79      | 11.29     |
| Howe         | 30.58      | 31.32   | 8.02       | 63.97    | 51.82    | 52.53   | 9.90       | 44.13     |
| Our proposed | 46.59      | 47.76   | 9.74       | 18.12    | 58.51    | 59.83   | 10.86      | 17.40     |

good enough. On the next sub section, the impact of these binarized images will be further discussed.

4.2. Experiment II -Segmentation Process

This sub-section discusses text line segmentation using true color image and binary image as input image. Then we manually evaluate the sum of the lines for each manuscript, as shown in Table 3. For this experiment, we use a number of input images like ground truth images (I_GT), Niblack’s binarized images (I_BN), Sauvola’s binarized image (I_BS), our proposed’s binarized image (I_BP), and original true color image (I_TC). Each lontar leaf on Kropak 22 consists of four lines. First, we do line segmentation of the ground truth images. The result is shown that the number of lines from ground truth images is as same as the number of the target line. Niblack’s binarized image produces five lines on six lontars, while the original true color image gained five lines on four lontars. It explains that when the quality of binarized image is poor, a better solution is to use the original image. For our proposed binarized image, the evaluation showed that the number of lines is the same as the target line number. Segmentation evaluation for Kropak 18 shows that there is no segmentation method that can perfectly reach the number of the target lines, as shown in Table 3. For document of 3306.tif, line segmentation that used a binarized image from Niblack and Sauvola generated the number of lines which are different from the target line number. Line segmentation using true color image as input produces fewer errors on the number of rows, while line segmentation of our proposed binarized image can produce nearly the same number of rows as the target.

5. Experiments and Result

We have built line segmentation framework that consists of three parts. Although the binarization process cannot entirely solve all the challenges of old Sundanese manuscripts, our proposed framework succeeded in improving the Niblack binarization method. We built a novel filtering to eliminate the groups of little noise so the binarized image was cleaner than the previous one. Hence, line segmentation for our proposed binarized image can closely achieve the target lines number.

Acknowledgments

We thank to Universitas Padjadjaran, for funding this research under the Internal Fundamental Research grant. We also thank to the Sundanese Culture Study and Kabuyutan Ciburuy for supplying the Sundanese old manuscripts. The authors is also deeply thankful to all suggestions given by Prof. Jean-Christophe Burie on old manuscript image analysis research.
Table 3. The comparison of the number of detected line for Kropak 18 and Kropak 22

| Images       | Target Line | (I_GT) | (I_BS) | (I_BN) | (I_BP) | (I_TC) |
|--------------|-------------|--------|--------|--------|--------|--------|
| CB-3-22-90-1 | 4           | 4      | 4      | 4      | 4      | 4      |
| CB-3-22-90-4 | 4           | 4      | 4      | 5      | 4      | 5      |
| CB-3-22-90-5 | 4           | 4      | 4      | 5      | 4      | 4      |
| CB-3-22-90-23| 4           | 4      | 5      | 4      | 4      | 4      |
| CB-3-22-90-33| 4           | 4      | 4      | 5      | 4      | 4      |
| CB-3-22-90-39| 4           | 4      | 4      | 5      | 4      | 4      |
| CB-3-22-90-40| 4           | 4      | 4      | 5      | 4      | 5      |
| CB-3-22-90-61| 4           | 4      | 4      | 4      | 5      | 4      |
| CB-3-22-90-62| 4           | 4      | 4      | 4      | 4      | 4      |
| 3306         |             | 4      | 4      | 6      | 1      | 4      |
| 3308         |             | 4      | 4      | 6      | 5      | 5      |
| 3309         |             | 4      | 4      | 6      | 6      | 4      |
| 3310         |             | 4      | 4      | 5      | 6      | 4      |
| 3311         |             | 4      | 4      | 5      | 6      | 4      |
| 3312         |             | 4      | 4      | 4      | 5      | 4      |
| 3313         |             | 4      | 4      | 4      | 5      | 4      |
| 3316         |             | 4      | 4      | 4      | 6      | 4      |
| 3322         |             | 4      | 4      | 5      | 4      | 4      |
| 3323         |             | 4      | 4      | 5      | 5      | 4      |
| 3349         |             | 4      | 4      | 5      | 1      | 4      |
| 3372         |             | 2      | 2      | 4      | 1      | 4      |

References

[1] Gatos B, Pratikakis I, and Perantonis S J 2006 Pattern Recognit. 39 317-327
[2] Kesiman M W A et al. 2017 Int. Jour. of Electronic Imaging 26 1-15
[3] Hadjadj Z, Cheriet M, Meziane A and Cherfa Y 2017 Signal, Image and Video Process 1-8.
[4] Khurshid K, Siddiqi I, Faure C and Vincent N 2009 SPIE Proc. 7247 1-9
[5] Sauvola J and PietikÄ„inen M 2000 Pattern Recognit. 33 225-236
[6] Ntolas N and Veintzas D 2008 Proc. 12th WSEAS Int. Conf. Commun. 41-51
[7] Chanchong R, Fung C C, Wong K W 2014 Cultural Comp 55-64
[8] Hassanien A E, Elfattah M A, Aboulenin S and Schaefer G 2016 IEEE Int. Conf. on Systems, Man, and Cybernetics 3842-3846
[9] Chen Y and Wang L 2017 Neurocomputing 237 272-280
[10] Pratikakis I, Gatos B and Ntirogiannis K 2013 Proc. Int. Conf. Doc. Anal. Recognit. 1471-1476
[11] Su B, Lu S and Tan C L 2013 IEEE Trans. Image Process. 22 1408-1417
[12] Louloudis G, Gatos B, Pratikakis I and Halatsis C 2008 Pattern Recognit. 41 3758-3772
[13] Mamatha H R and Srikantamurthy K 2012 Int. J. Appl. Inf. Syst. 4 13-19
[14] Bar-Yosef I and Hagbi N 2009 Proc. Int. Conf. Doc. Anal. Recognit. 1161-1165
[15] Howe N R 2013 Int. J. Document Anal. Recognit. 16 247-258
[16] Arvanitopoulos N and SÄijstrunk S 2014 Int. Conf. Front. Handwrit. Recognit. 726-731
[17] Garz A, Fischer A, Sablatanig R and Bunke H 2012 Proc. - 10th IAPR Int. Work. Doc. Anal. Syst. 95-99
[18] Boiangiu C A, Ioanitescu R and Tanase M C 2014 Int. J. Comput. Commun. Control 9 8-15
[19] Saund E, Lin J and Sarkar P 2009 Proc. Int. Conf. Doc. Anal. Recognit. 646-650
[20] Clausner C, Pletschacher S and Antonacopoulos A 2011 Proc. Int. Conf. Doc. Anal. Recognit. 48-52
[21] Liao W, Dong N and Fan T 2009 IEEE Int. Works. on Open-source Soft. for Scient. Comput. 88-90