J. Sains Dasar 2017 6 (2) 133 - 142

BINARIZATION AND SEGMENTATION FRAMEWORK FOR SUNDANESE ANCIENT DOCUMENTS

Erick Paulus[1]* , Mira Suryani[2], Setiawan Hadi[3], Rahmat Sopian[4], Akik Hidayat[5]

[1,2,3,5] *Department of Computer Science, UniversitasPadjadjaran, Indonesia*
[4]*Sundanese Culture Studie, UniversitasPadjadjaran, Indonesia*

*email: erick.paulus@unpad.ac.id

Abstract

Binarization and segmentation process are two first important methods for optical character recognition system. For ancient document image which is written by human, binarization process remains a major challenge.In general, it is occurring because the image quality is badly degraded image and has various different noises in the non-text area.After binarization process, segmentation based on line is conducted in separate text-line from the others. We proposedanovel framework of binarization and segmentation process that enhance the performance of Niblackbinarization method and implementthe minimum of energy function to find the path of the separator line between two text-line.For experiments, we use the 22 images that come from the Sundanese ancient documents on Kropak 18 and Kropak22. The evaluation matrix show that our proposed binarization succeeded to improve F-measure 20%for Kropak 22 and 50% for Kropak 18 from original Niblack method.Then, we present the influence of various input images both true color and binary image to text-line segmentation. In line segmentation process, binarized image from our proposed framework can producethe number of line-text as same as the number of target lines. Overall, our proposed framework produce promised results so it can be used as input images for the next OCR process.

Keywords: binarization, segmentation, ancient document

Introduction

The discipline of Computer Sciences is closely related to mathematics[1][2]. Mathematics fields such as discrete mathematics, linear algebra and logic has given a big impact for improvement of image processing methodologies especially in optical character recognition (OCR). As we know that binarization and segmentation process are two crucial process of OCR system. In both process, mathematicsmodel and logical problem solvingis needed to solve the challenges of image processing and image analysis.

The important step in document image processing is binarization process[3]. The main goal of binarization process is to convert gray scale image into binary image[4]. So, we can analysis the image content such as the number of lines, words and characters. In fact, the ancient document images commonly have low contrast, random noises, non-uniform illumination and fading, so binarization is still real and open challenges [4][5].Many paper explain that line segmentation process need binarization process first[6]. However, there are some researches discuss the opportunities in segmentation method that can extract line-text from binarized image or original image [7][8]. Text-line segmentation is the early step in text extraction procedure. After that, we can continue to extract word, syllable, or character. In this paper, we only discuss and evaluate our proposed framework from binarization process until line segmentation process.

The work of Binarization process is to compute the threshold value based on global or local. However, many studies showed that local thresholding is better for ancient document[9].

This paper is conducted as follow: Section II presents a description about the collection of Sundanese ancient document and the challenges for binarization and text line segmentation. Section III gives the detail description about the proposed framework. The experiments and result of the proposed framework is explained in Section IV. Last Section describes the conclusions and some opportunities for the future study.

SUNDANESE DOCUMENT

The Collection of Sundanese Ancient Documents

In West Java, most of the collection of Sundanese ancient documents stored in museums. But, there are some other places that keep the

original and the duplicate of the documents. One of the place that keep Sundanese ancient documents in form of its original is KabuyutanCiburuy, Garut, West Java, Indonesia. The place keep the Sundanese ancient documents written on palm leaf called Sundanese Lontar. We collaborate with the owner of the documents and use plam leaf documents as dataset for this study. The Sundanese Lontar is about 25 to 45 cm and the width is about 10 to 15 cm. There are 27 collections stored in wooden box. Each collection consists about 25 to 30 pages vice versa. Each pages consists of 4 to 6 lines, and estimated there are 15 to 20 words in each line. The Sundanese Lontar has hole in the center and bind by the rope. The various content such as Ramayana story, farm formulae, code law, medicine formulae, and life message made the Sundanese Lontar extremely valuable for Sundanese. The content of the Sundanese lontar is Sundanese ancient script and written by using 'Peso Pangot'. The writing process of the Sundanese Lontar is stopped, but, there are some philologists can read and transliteration the documents. The samples of Sundanese ancient manuscripts is shown on Figure1.

The Challenges for Binarization and Line Segmentation

Due to aging and the degraded condition of the storage, the Sundanese Lontar are in poor quality. The manuscripts became moldy, have many fractures, non-uniform illumination, and many random noises. Some of the Sundanese Lontar even split into two or more parts. Moreover, the physical condition of the manuscripts have been experiencing brittleness and discoloration such as smear, spot, stain, scratch, and shadow , as shown on Figure 2. Due to the lack of knowledge, it create mismatch part when recombined. All of the characteristics are the challenges for every step in image processing include binarizarion and text line segmentation.

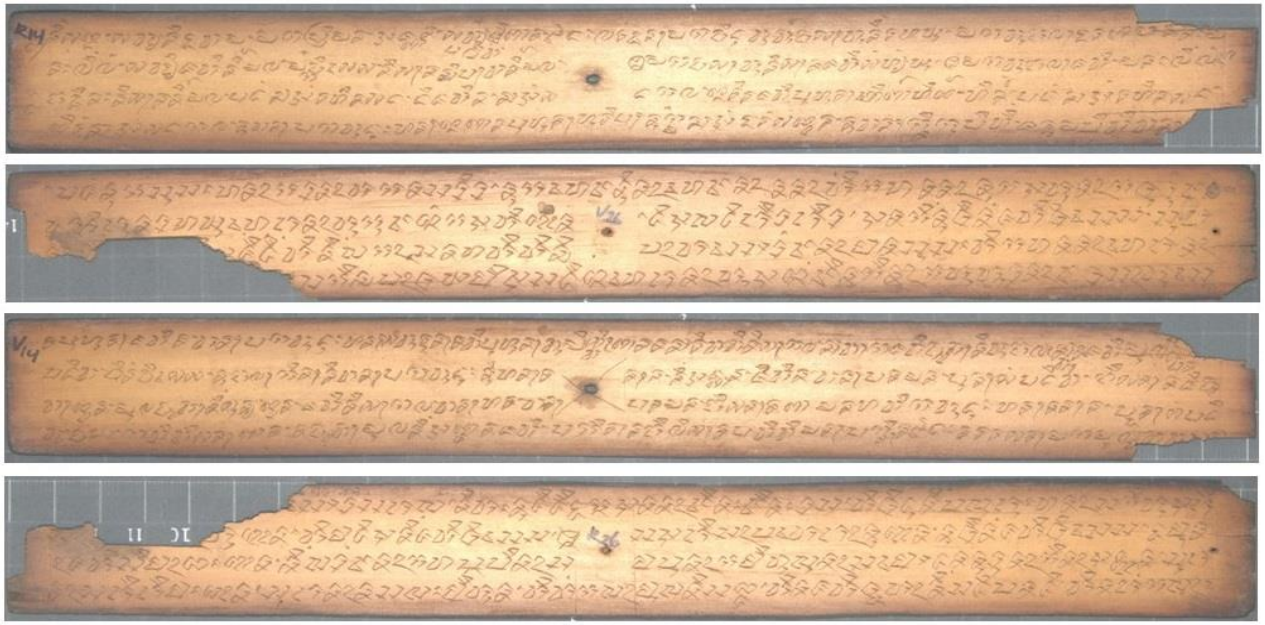

Figure1. The samples of Sundanese ancient manuscripts on palm leaf.

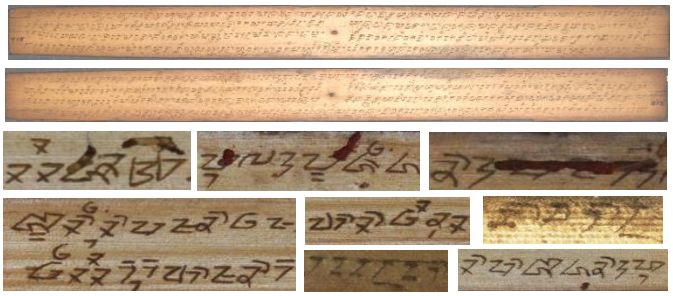

Figure 2. Various physical condition of the ancient Sundanese manuscripts.

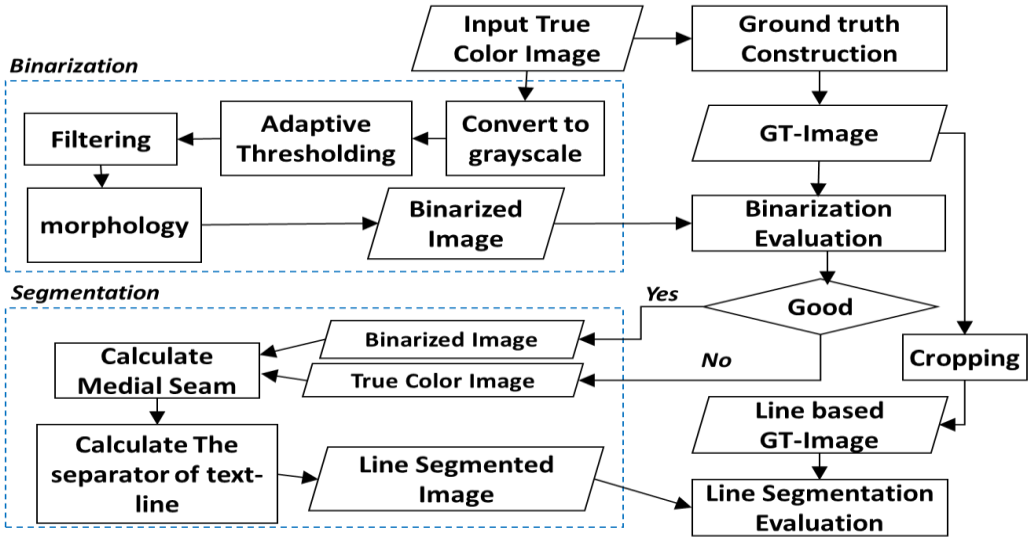

FIGURE 3. The proposed framework of binarization and line segmentation

Observation And Analyzation

1. The Proposed Framework

We design our proposed framework to evaluate the performance of binarization and line segmentation methods, as shown on Figure 3. This framework consist of 3 part processes. The first process is to generate ground truth image. We need the ground truth image to compare the binarized image using evaluation matrix. The second process is to get the optimal binarized image using our proposed method. We proposed a novel filtering to increase the performance of Niblackbinarization. Then, the third processpresents line segmentation. This framework is design for our Sundanese ancient documents.

2. Binarization Process

An adaptive binarization, filtering and morphologymethod is implemented to the original image to make sure the optimal binarizationresult for the next step OCR. The algorithm ofbinarization is described as:

1. Convert the original image from RGB to grayscale.
2. Convert the grayscale image into binary image using local based thresholding.
   Conversion digital image into binary image is determined by *T* threshold value. For this study, we implement some thresholding method such as Otsu[10][11], Niblack[12], Sauvola[13], Howe[14].The binarization processcan be formulated on equation (1).

$$bw(i,j) = \begin{cases} 1\,(white), & if\ \ f(i,j) \geq T \\ 0\,(black), & otherwise \end{cases} \quad (1)$$

Where *bw* is binary image that contain only white or black.

Some cases in ancient documents need more treatment such as adaptive binarization using local thresholding.Niblack, an adaptive binarization, calculate the local threshold (*T*) values within a certain sub window (mask) of the pixel size N x N. Equation (2) present the Niblackbinarization formula.

$$T = m + k * s \quad (2)$$

where*m* is the average and *s* is the standard deviation of color intensity I in the sub window, and *k* is a constanta value in [-1,0). *k* = -0.2 is good for black object detection and *k*= + 0.2 is good for white object detection.This method remain artifact and noises for areas with badly degraded image.

3. Filtering

   Filtering is used to remove noise that distributed in digital image area. We proposed a novel filtering to enhance Niblack method, as shown on equation (3). The idea is to erase a small group noise, that containing black area is less than a half of white area.

$$C(i,j) = \sum_{m=-r}^{r} \sum_{n=-r}^{r} f(i+m, j+n)$$

$$bw_{filter}(i,j) = \begin{cases} 1\,(white), & if\ C(i,j) \geq \lfloor (2*r+1)^2/2 \rfloor \\ 0\,(black), & otherwise \end{cases} \quad (3)$$

Where r is a half-length (N) of sub window (N x N), C is cumulative sum of white area

4. Remove circle object in the middle documents
   Generally,Sundanese ancient document is stored in the box that contain some palm leaf manuscripts(Lontar). At the center of the lontar there is a small hole to insert the rope as the bond. When the data acquisition process, this hole is recorded as black and needs to be removed in order to produce a better line segmentation. Closing morphology with a disk-shaped is used to in this removing method. Illustration of closing disk morphology can be seen in Figure 4.

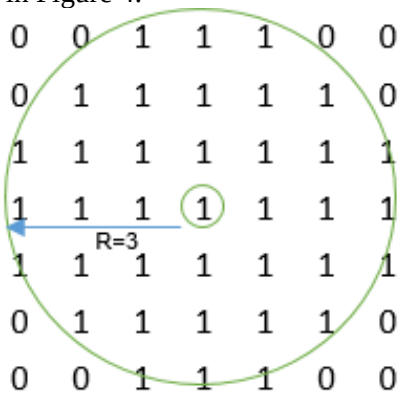

FIGURE 4. Morphological closing with shape-disk

3. Segmentation Process

Projection profile scheme is one of the common basic method for line segmentation[15][16][17]. There are two type orientation of projection profile , namely the horizontal projection profile and the vertical projection profile. Horizontal projection profile is commonly used for line segmentation and vertical projection profiles can be used for segmentation of words or characters.

In this study, Profile Projection Horizontal (PPH) is used as a representation of a histogram of black pixels summation that accumulated binary image along parallel lines in a document. The mathematical functions of PPH can be seen in equation (4). Image pixels defined by the function *f (x, y)* where *x* and *y* respectively represent the row and column. The parameter n describes the number of columns in an image

$$PPH(x) = \sum_{1 \leq y \leq n} f(x,y) \quad (4)$$

The algorithm of text line segmentation is adopted from seam carving method [18][7][19]. Some studies explain that this algorithm is free from binarization. Hence, it can be an alternative solution for our ancient documents segmentation. The following is a segmentation algorithm :

A. Compute medial seam

Calculations along the midpoint of the line is based on the following sequence:

1. Image notation
   $I \in \Re_{a\,x\,b}$ (5)

   With*a* :row and*b*: column
   N=a; m=b, r=p, g=h
2. Divide the page vertically into *p* pieces, with w is width

   $$w = \lfloor b/p \rfloor \quad (6)$$

3. Edge Detection using Sobel Operator
4. Calculate the horizontal projection smoothed profile for each piece P_h^c

   $$P_i^c = \sum_{j=k}^{k+w-1} S_{i,j},\ P^c = \{P_i^c\}_{i=1}^{n},\ P_h^c = h(P^c) \quad (7)$$

   $c = 1, \ldots, r,\ k \in \{1, 1+w, \ldots, 1+(r-1)w\}$
   with h is a cubic spline smoothing filter

5. Find and connect the nearest local maximum of each piece

B. Compute the separator of text-line

The calculation of separating lines implement modified Seam Carving procedure that described as follows

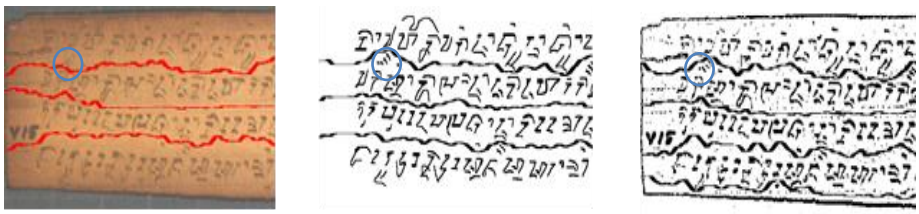

Figure5. Example of Line segmented image for true color image and binary image (left to right)

1. Calculate Energy Function :

$$E_{i,j} = \left| \frac{I^{\sigma}_{i,j+1} - I^{\sigma}_{i,j-1}}{2} \right| + \left| \frac{I^{\sigma}_{i+1,j} - I^{\sigma}_{i-1,j}}{2} \right| \quad (8)$$

with $I_{\sigma}$is the grayscale image that smoothed by Gaussian filter for standard deviation σ

2. Calculate the cumulative of minimum energy M

$$M_{yh(j),1} = E_{yh(j),1}, \quad (9)$$

$$M_{yh(j),j} = E_{yh(j),j} + \min \begin{cases} M_{yh(j)-1,j-1} \\ M_{yh(j),j-1} \\ M_{yh(j)+1,j-1} \end{cases}$$

3. Calculate the optimal path from cumulative energy M. The high-energy regions represent text component and low-energy regions represent the background.

The sample imageof separating text-line is shown in Figure 5. Sometime there are some component texts which is not include in their lines.

Results And Discussion

There are 22 Sundanese ancient document that consist of 12 true color images fromKropak18 and 10true color images from Kropak 22. In general, every image contain 4 lines. For ground truth images, we manually construct the binary ground truth imagesusing PixLabelers[20] tools. Then, wesegmented it become the text-line ground truth images using Alethea Lite[21] tools and build an application program of image cropping based on XML file using Scilab[22]. The diagram of ground truth image construction is shown in figure 2.

There are some evaluations that are designed to measure the performance of the binarization method of document image. Pratikakis[10] mentioned that there are three main categories of evaluation methods, namely the evaluation of visual inspection by one or more testers (humans), the evaluation is intended to measure the performance of OCR, and evaluation of the comparison pixel to pixel between the binarized image and ground-truth image. Evaluation-based pixel is the most widely method used for measuring the performance of the binarization method of document image. This work uses pixel-based evaluation to measure binarization method, such as

- F-Measure (F-Measure), The higher of F-Measure value, the more accurately binarized image
- Fps-Measure (Pseudo F-Measure), The higher of pseudo F-Measure value, the more accurately binarized image
- PSNR (Peak signal to noise ratio), the higher of PSNR value show the binarized image is closer to the ground truth image. PSNR is good to evaluate the true color or grayscale image.

Experiment I - Binarization Process

The first sub framework is proposed to enhance the binarizationperformance for Sundanese document.Based on our experiment on the implementation of some common binarization methods, we find the advantage of Niblack method. Even though Niblack remain many noise on its binarized image, but the contour of text is visually more readable. Hence, the idea of our binarization framework is to remove as much as noise so we can increase the readability of text. We have compared our proposed framework to some common binarization methods likeOtsu[10],[11], Niblack[12], Sauvola[13], and Howe[14]. Figure 5 present the example of binarized images. Otsu's binarized image result many dark area like ellipse

shape on the corner image. For Niblack and Sauvola, the contour of text is still clear and better than other common methods. Howe's binarize image show that many text on the left and right side image is removed.After we do some filtering methods for Niblack methods, many noise that distributed in the middle and edge image has reduced. Based on experiment, our proposed binarization framework succeeded to get the highest F-measure, pF-measure, and PSNR than other methods, as shown in table 1-3. It means that our proposed method produce promised results.

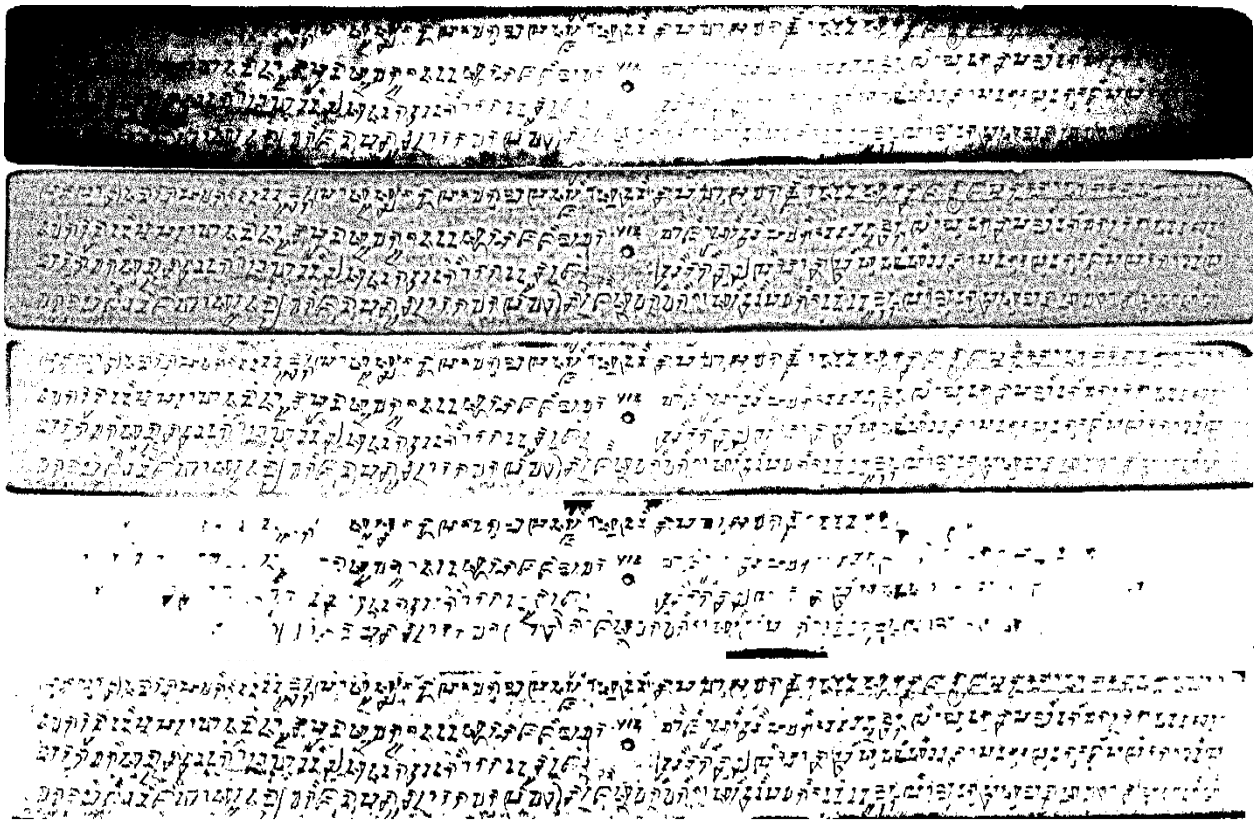

FIGURE 6. Example of binarized images(up to bottom)using method of Otsu[10],[11], Niblack[12], Sauvola[13], Howe[14], proposed scheme

TABLE 1. TheBinarizationevaluation matrix based on ICDAR competition[10] for the original image of 3322.tif

Ancient Doc Images	Precision	Recall	F-measure (%)	pFMeasure (%)	NRM	PSNR	DRD	MPM (x1000)
Otsu_3322.tif	0.67	0.14	23.15	24.70	0.46	3.86	37.96	181.88
niblack_3322.tif	0.22	0.67	32.81	33.51	0.29	5.98	46.42	144.77
sauvola_3322.tif	0.32	0.33	32.19	34.75	0.37	8.96	20.81	63.35
Howe_3322.tif	0.47	0.32	37.97	47.75	40.15	0.36	10.21	14.16
Our proposed _3322.tif	0.44	0.65	52.32	54.37	0.22	9.63	17.46	24.56

TABLE 2. Averagebinarization evaluation matrix based on ICDAR Competition[10] for Kropak 22

Ancient Doc Images	Precision	Recall	F-measure (%)	pFMeasure (%)	NRM	PSNR	DRD	MPM (x1000)
Otsu	0.15	0.82	26.03	26.21	0.29	4.11	70.90	304.41
Niblack	0.36	0.84	50.54	51.32	0.15	8.72	22.31	68.60
Sauvola	0.66	0.43	51.65	55.62	0.29	11.79	8.50	11.29
Howe	0.42	0.68	51.82	52.53	0.20	9.90	15.52	44.13
Our proposed	0.50	0.71	58.51	59.83	0.18	10.86	11.80	17.40

TABLE 3. Averageofbinarization evaluation matrix based on ICDAR Competition[10] for Kropak 18

Ancient Doc Images	Precision	Recall	F-measure (%)	pFMeasure (%)	NRM	PSNR	DRD	MPM (x1000)
Otsu	0.67	0.13	21.44	23.78	0.46	4.23	34.02	158.68
Niblack	0.19	0.66	29.12	29.56	0.30	5.78	49.59	146.18
Sauvola	0.26	0.39	30.82	32.30	0.35	8.88	23.53	64.46
Howe	0.24	0.48	30.58	31.32	0.33	8.02	30.22	63.97
Our proposed	0.34	0.65	44.11	45.18	0.23	9.11	22.64	31.45

Table 2 and table 3 present the evaluation matrixof various different binarization methods for ancient Sundanese documents.Table 2 describe the binarization performance for Keropak 22 using some common binarization methods. In general, Niblack, Sauvola, Howe dan our proposed method get the average of F-measure above 50%. But the highest value of F-measure is obtained by our proposed method, which amounted 58,51. While Sauvola acquire the highest PSNR about 11.79, followed by our proposed method about 10.86. The sauvola'sbinarized image is closer to ground truth image.

The F-measure average of Kropak 18 is still under 50% because the physical condition of documents have many smudge, black spot, and low contrast. Besides, the MPM average of Kropak 18 is decrease from 0.146 point into 0.031 point. It meansthat the binarized image is more closely to ground truth image but it does not good enough. Next sub section, we will discuss the impact of these binarized images.

Experiment II -Segmentation Process

This sub section is discussed about text line segmentation using the true color image and binary image as an input image. Then we manually evaluate the sum of the lines for each Sundanese documents, as shown on Table 4. For this experiment, we use some input images like groundtruth images($\boldsymbol{I_{GT}}$), Niblack'sbinarized images ($\boldsymbol{I_{BN}}$), Sauvola'sbinarized image ($\boldsymbol{I_{BS}}$), our proposed'sbinarized image ($\boldsymbol{I_{BP}}$), and original true color image($\boldsymbol{I_{TC}}$). Each lontar on Kropak 22 consist of 4 lines. First, we do line segmentation for the ground truth images. The result is shown that the number of lines from ground truth images is as same as the number of target line. Niblack'sbinarized image produce 5 lines on 6 lontars. While and original true color image gain 5 lines on 4 lontars. It means that when the quality of binarized image is not good, then using the original image is a better solution. For our proposed binarized image, the evaluation shown that the number of lines is as same as the number of target line.

Segmentation evaluation forKropak 18 shows that there is no segmentation method which perfectly achieve to the number of target lines.

TABLE 4. Thecomparison of segmentation evaluation for Kropak 22

Ancient Document Images	Number of Target Line	Number of Detected Line (I_{GT})	Number of Detected Line (I_{BS})	Number of Detected Line (I_{BN})	Number of Detected Line (I_{BP})	Number of Detected Line (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	5
CB-3-22-90-30	4	4	4	5	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	4	5
CB-3-22-90-62	4	4	4	4	4	4
3306	4	4	6	1	4	5
3308	4	4	6	6	5	5
3309	4	4	6	6	4	4
3310	4	4	5	6	4	3
3311	4	4	5	5	4	4
3312	4	4	4	5	4	4
3313	4	4	5	5	4	4
3316	4	4	4	6	4	3
3322	4	4	5	4	4	4
3323	4	4	5	5	4	4
3349	4	4	5	1	4	5
3372	2	2	4	1	4	4

I_{GT} = ground truth image; I_{BN} = Niblack'sbinarized image;I_{BS} = Sauvola'sbinarized image; I_{BP}= ourproposed'sbinarized image;I_{TC}= true color image

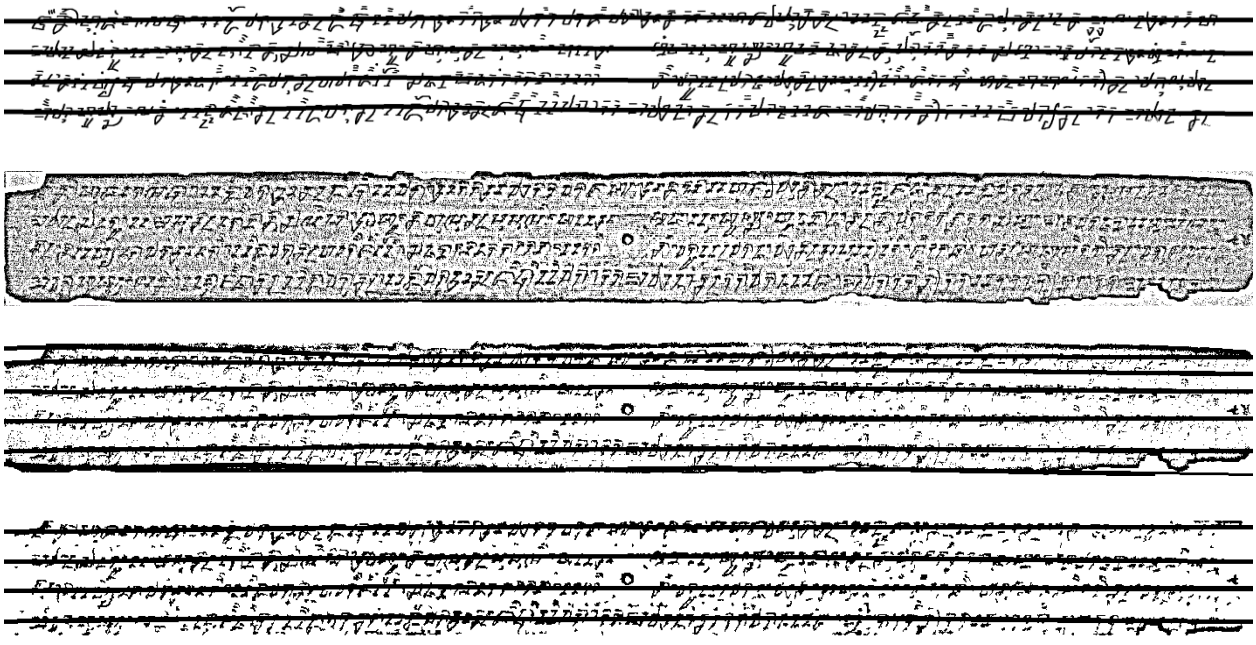

FIGURE 7. Example of line segmentation results for image 3306.tiff using input images (up to bottom) from ground truth, Niblack, Sauvola, our proposed method, true color

Conclution

There is no binarization scheme that able to complete all the challenges of ancient Sundanese manuscripts. But, our proposed framework is succeed to improve the Niblackbinarization method. We build a novel filtering to eliminate the groups of little noise so the binarized image is cleaner than before. On the next process, line segmentation for our proposed binarized image can closely achieve the number of target lines. For Sundanese documents, binarization process is still open real challenges. So we can improve the binarization performance for the next researches using other method like graph theory.

Acknowlegment

Wegive thank to Kabuyutan Ciburuy in West Java and philologist from Sunda Departement, Unpad, for providing us the sample of ancient sundanese manuscripts. We thank to UniveristasPadjadjaran for continuing support in the Internal Fundamental Research. Early, this work is supported by the International Research Collaboration And Scientific Publication Program implemented by the Ministry of Research, Technology and Higher Education Indonesia. Also we give great appriciation to Prof. Jean-Christophe for collaboration in AMADI Project from the University of La Rochelle, France.

References

[1] D. Baldwin, H. M. Walker, and P. B. Henderson, "The roles of mathematics in computer science," *ACM Inroads*, vol. 4, no. 4, pp. 74–80, 2013.

[2] P. B. Henderson, "The Role of Mathematics in Computer Science and Software Engineering Education," *Adv. Comput.*, vol. 65, no. 5, pp. 349–395, 2005.

[3] Z. Hadjadj, M. Cheriet, A. Meziane, and Y. Cherfa, "A new efficient binarization method: application to degraded historical document images," *Signal, Image Video Process.*, 2017.

[4] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognit.*, vol. 39, no. 3, pp. 317–327, 2006.

[5] M. W. A. Kesiman, S. Prum, J. C. Burie, and J. M. Ogier, "An initial study on the construction of ground truth binarized images of ancient palm leaf manuscripts," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 656–660.

[6] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "Text line detection in handwritten documents," *Pattern Recognit.*, vol. 41, no. 12, pp. 3758–3772, 2008.

[7] A. Garz, A. Fischer, R. Sablatnig, and H. Bunke, "Binarization-free text line segmentation for historical documents based on interest point clustering," *Proc. - 10th IAPR Int. Work. Doc. Anal. Syst. DAS 2012*, pp. 95–99, 2012.

[8] A. Garz, A. Fischer, H. Bunke, and R. Ingold, "A binarization-free clustering approach to segment curved text lines in historical manuscripts," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 1290–1294, 2013.

[9] B. Su, S. Lu, and C. L. Tan, "Robust document image binarization technique for degraded document images," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1408–1417, 2013.

[10] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2013 document image binarization contest (DIBCO 2013)," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, no. Dibco, pp. 1471–1476, 2013.

[11] N. Ntogas and D. Ventzas, "A Binarization Algorithm For Historical Manuscripts," in *12th WSEAS International Conference on communications*, 2008, pp. 41–51.

[12] K. Khurshid, I. Siddiqi, C. Faure, and N. Vincent, "Comparison of Niblack inspired binarization methods for ancient documents," *SPIE Proc.*, vol. 7247, p. 72470U–72470U–9,

2009.
[13] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognit.*, vol. 33, no. 2, pp. 225–236, 2000.
[14] N. R. Howe, "Document Binarization with Automatic Parameter Tuning," *Int. J. Documant Anal. Recognit.*, vol. 16, no. 3, pp. 247–258, 2013.
[15] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *Int. J. Doc. Anal. Recognit.*, vol. 9, no. 2–4, pp. 123–138, 2006.
[16] R. P. Dos Santos, G. S. Clemente, T. I. Ren, and G. D. C. Calvalcanti, "Text line segmentation based on morphology and histogram projection," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 651–655, 2009.
[17] H. R. Mamatha and K. Srikantamurthy, "Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document," *Int. J. Appl. Inf. Syst.*, vol. 4, no. 5, pp. 13–19, 2012.
[18] N. Arvanitopoulos and S. Süsstrunk, "Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts," *Int. Conf. Front. Handwrit. Recognit.*, no. Ic, pp. 726–731, 2014.
[19] C. A. Boiangiu, R. Ioanitescu, and M. C. Tanase, "Handwritten documents text line segmentation based on information energy," *Int. J. Comput. Commun. Control*, vol. 9, no. 1, pp. 8–15, 2014.
[20] E. Saund, J. Lin, and P. Sarkar, "PixLabeler: User Interface for Pixel-Level Labeling of Elements in Document Images," in *2009 10th International Conference on Document Analysis and Recognition*, 2009, pp. 646–650.
[21] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments," in *2011 International Conference on Document Analysis and Recognition*, 2011, pp. 48–52.
[22] W. Liao, N. Dong, and T. Fan, "Application of Scilab in teaching of engineering numerical computations," in *2009 IEEE International Workshop on Open-source Software for Scientific Computation (OSSC)*, 2009, pp. 88–90.