Javanese Character Image Segmentation of Document Image of *Hamong Tani*

Agustinus Rudatyo Himamunanto
Faculty of Sciences and Computer
Immanuel Christian University
Yogyakarta, Indonesia
rudatyo@ukrimuniversity.ac.id

Anastasia Rita Widiarti
Faculty of Sciences and Engineering
Sanata Dharma University
Yogyakarta, Indonesia
rita_widiarti@usd.ac.id

***Abstract*—Script image segmentation of a document image is the most decisive step to the success of the process of transliteration of the script image into another script, such as automatically transliterating a printed Javanese manuscript image into a Latin manuscript. This paper gives an example of the application of profile projection modification to the segmentation of Javanese script document image of the entire 87 pages of the document image of *HamongTani* book. Based on the output of the developed system, the average percentage of correctness is 84.255% with the average standard deviation of 14.093%. This value of average percentage of correctness shows that the model developed for the Java script document image segmentation of the *HamongTani* book is relatively good.**

Keywords—Hamong Tani; profile projection modification; character segmentation

I. INTRODUCTION

Yogyakarta, Indonesia, is an area known for its rich history of culture, including art, and literature. One of the history's invaluable wealth is the Javanese manuscripts that are often found in Yogyakarta, namely in the Keraton Kasultanan, and Pura Pakualaman [1]. Digitalizing these texts presents many benefits such as the improvement of their quality, the extension of their existence, and the capability to be manipulated for research purposes such as being automatically transliterated into Latin script without changing the meaning of the original text.

Hamong Tani is one of many culture heritages in Javanese literature. This book was written by Karel Frederik Holle in Dutch language and translated to Javanese languge in Javanese script by F. L. Winter in 1894. *Hamong Tani* is an important book because it marked the beginning of the entry of the Dutch agricultural technology into traditional Javanese society. *Hamong Tani* teaches better ways of farming to the Javanese community at the time. This book will be very valuable when transliterated into Latin because Indonesia is an agricultural country even now. Manual transliteration takes a very long time and can only be done by a philologist. Research in the field of automatic transliteration from the Javanese script into theLatin would be very useful.

The development of the science of document image analysis, namely the analysis of the visual representation of paper documents such as journals, the facsimile, office documents, spreadsheet, and so forth [2] opens up great opportunities to be exploited for the preservation of many ancient texts found in Yogyakarta. O'Gorman and Kasturi [3] describes the stages of the process of document image analysis which can be modified to document image recognition of Javanese literatures. It begins with data acquisition in which data from paper documents is read by an optical scanning device and the results are saved as a digital image file. The next stage is pixel-level processing stage aimed at preparing the document image, and making intermediary features to help with image identification. The third stage is the stage of character recognition which aims to translate a string of characters with a variety of shapes and sizes.

Research in an automatic digital transliteration is divided into 3 main steps:

a. segmentation process of the image of Javanese characters from the image document,

b. transliteration of Javanese characters from the segmentation results, and

c. grouping of the transliteration results into syllables.

This paper presents the results of research in the segmentation of Javanese characters as the first step in the automatic transliteration process.

The main purpose of the segmentation process is to automatically extract each character image contained in a text document image. This extracted character will be processed in the next stage. Some advantages of a successful segmentation process are:

a. The failure rate in the process of character recognition can be reduced.

b. The relatively simple form of data will reduce memory usage.

c. The process in the next stage will be faster because the stored data is relatively simple and more efficient.

978-1-4799-3170-5/13/$31.00 ©2013 IEEE

II. Study Of The Characteristics And Rules Of Writing Javanese Script

Javanese scripts consist of two major scrip groups, which are basic Javanese scripts and derivative Javanese scripts. Basic Javanese scripts are main Javanese scripts which haven't been added with various punctuation marks or *sandangan*, therefore these main Javanese scripts are called *legena* or *wuda* scripts which means bare scripts. Fig. 3 shows 20 *legena* Javanese scripts.

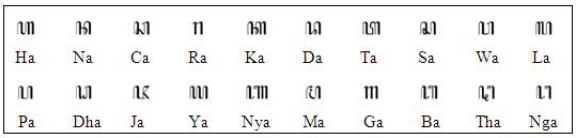

Fig. 3. *Nglegena* Javanese scripts

Generally, most Javanese scripts used don't only consist of *legena* scripts, but use various additions *sandangan*. There are many kinds of *sandangan*, i.e. *sandangan swara* for i consonant called *wulu* and is written above corresponding *legena* script. Or *sandangan swara* of u consonant called *suku* which is written beneath *legena* script.

From a study of placement area of Javanese scripts and punctuation marks or *sandangan,* an information on spots to place Javanese scripts is obtained, as shown in Fig. 4.

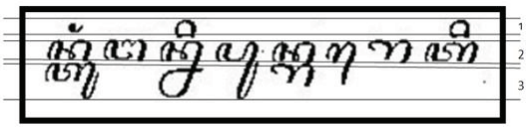

Fig. 4. Writing area of Javanese script

Area 1 is upper area or upper zone. In this zone there are *sandangan swara* i.e. *wulu* and *pepet,* and *sandangan panyigeg* i.e. *layar* and *cecak*. Area 2 is main area or main zone. It's called main zone because this is where basic scripts of Javanese scripts which are *legena* scripts. Area 3 is lower are or lower zone. Lower zone is a place for *suku*, lower part of *taling*, *cakra mandaswara* script, *cakra keret*, as well as *pasangan* placed below. Fig. 5. shows examples of Javanese scripts and places to put the scripts.

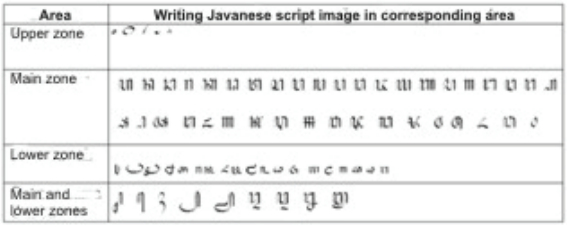

Fig. 5. Writing zones of Javanese scripts

Fig. 6 shows the practice of Javanese script writing, Javanese scripts are written from left to right, and one Javanese script can take the form of:

1) *Legena* script only, i.e. *sa* script in Fig. 6(a).
2) *Legena* script with *sandangan* placed on the upper zone, i.e *se* script in Fig. 6(b).
3) *Legena* script with *sandangan* placed on the lower zone, i.e. *su* script in Fig. 6(c).
4) *Legena* script with *sandangan* in upper and lower zones at once, i.e. *sur* script in Fig. 6(d).

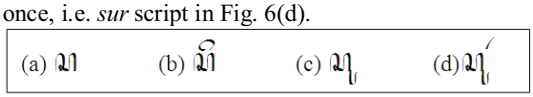

Fig. 6. Samples of script forms with combination of the placement of the additions

III. Image Segmentation

Image segmentation is the process of breaking an image into objects that are contained in it [4]. In document image analysis, the segmentation process is divided into two steps. The first step will separate the text images from figure images, and the second step will do a further separation process of the result from the first step [5] [6] [7]. For example, the next process for text images is to transcribe the text into its components, that is to find the columns, paragraphs, words, until the characters that form the word are finally found. The result of this segmentation process is then used for further image processing such as pattern recognition.

There are several methods that can be used in image segmentation, one of which is the projection profile method. Zramdini [7] formulates, that if there is a binary image S with M rows and N columns, then the vertical projection profile (P_v) of the image S is the number of black pixels perpendicular to the y-axis, i.e.

$$P_v[i]=\sum_{j=1}^{M} S[i,j] \quad (1)$$

while the horizontal projection profile (P_h) of the image S is the number of black pixels perpendicular to the x axis, namely

$$P_h[i]=\sum_{j=1}^{N} S[i,j] \quad (2)$$

IV. Preprocessing

Input in this research is the scanned images from each page of *Hamong Tani* using imaging devices that have a resolution of 300 dpi. Prepocessing is conducted on the results of the scan so that the next process can be performed optimally. Some preprocessing steps are conducted on the input image, among others: binarization, noise reduction, skew detection and the document image rotation. The binarization process is done so that information of the input image becomes simpler as the automatic transliteration process does not require color or degrees of gray. The binarization process in this research uses im2bw(), a function of Matlab.

A scanned image may also have many types of salt and pepper noise. A group of pixels is categorized as noise if it consists of one to three pixels. This size is determined based on some observations.

Skew detection is necessary because the sheet being scanned can not be perfectly straight. Damage may happen to the page if it is straightened by force. On the other hand, the segmentation process which uses the profile projection method will not work well. Skew detection method used in this

research is the moment orientation method. The angle values obtained will be used to rotate the slope of the relevant page. The rotation process in this research uses Matlab's imrotate() function.

V. Model of Image Segmentation of Documents with Javanese Script

Projection profile has opened many opportunities to obtain the images of characters that form a document image, especially if there is a clear distance between rows and between characters as in Fig. 1.

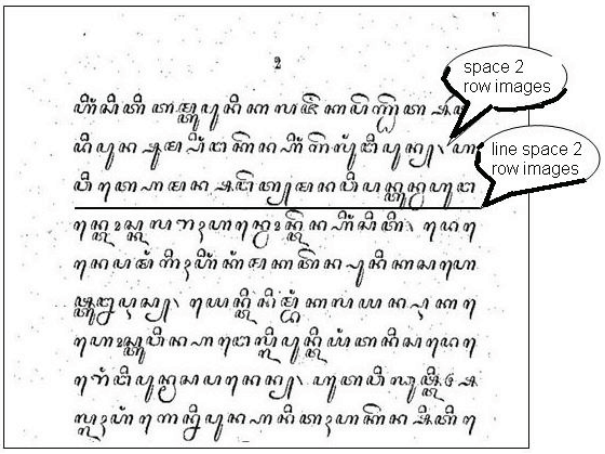

Fig. 7. Visualization of space between 2 rows and line image to separate 2 rows

Taking the characteristics of printed Javanese script document image and the workings of the profile projection method into account, the method that was developed in this research to find the segmentation of the character images from the *HamongTani* book is as follows:

1. Create horizontal projections to find the position of the row's bottom and row's top for each row in the document images.
2. Cut the rows of character images based on index position of rows, which is the result of the horizontal projection operation.
3. For each row of character images that has been found, vertical projection is then performed to find the column index of each character on the row.
4. Cut the characters from the appropriate character row using column index generated by the operation of the vertical projection on the line.

The method above is then tested on the entire manuscript image of the *Hamong Tani* book totaling 87 pages.

To test the performance of this method, a technique is designed to determine the percentage of correctness. The simplest way to determine the percentage of correctness is to count the correct segmentation results with the help of Java literature experts. All the correct data is then divided with all the data being tested. Equation (3) below will calculate the correctness rate:

$$percentage\ of\ correctness = \frac{\sum correct\ \ data}{\sum data\ \ being\ \ tested} x\ 100 \qquad (3)$$

VI. Results and Discussion

The testing of the image segmentation of Javanese script image system was applied on the scanned text document image of the *Hamong Tani* book. *Hamong Tani* was chosen because of its ancientness, good physical condition and completeness. Moreover, this book contains the ancient agricultural knowledge that is still relevant to the current agricultural and moral knowledge that will help young people to appreciate the services of farmers.

After all the pages of *Hamong Tani* are scanned, then the process of preprocessing, i.e. binarization, noise removal, skew detection, and page rotation is performed. The next process is segmentation algorithm using a modified projection profile.

Fig. 2 shows the visualization of the character image segmentation results of page 16 of the *Hamong Tani* book. Also shown are the character segmentation result of rows 1 through 13 and the image line numbers.

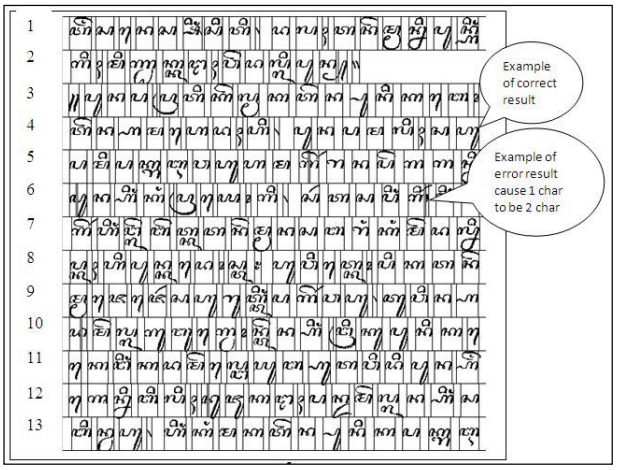

Fig. 8. Visualization of the Javanese character image segmentation results

The number of correct segmentation results on each row and the total number of correct segmentation results on the page can be counted manually from the segmentation result of page 16 of *Hamong Tani*. In the first line it can be seen that there are 18 Javanese characters and 18 characters in a row according to the segmentation result, so it can be said that all the characters on the first line are segmented correctly. The same steps are done on the next rows until it is finally known that on page 16 there are 328 correct character segmentation results of the 337 characters that should be found.

The percentage of correctness of segmentation results page 16 can be calculated using formula (3) as follows:

$$persentage\_of\_correctnes = \frac{\sum correct\_data}{\sum data\_being\_tested} x100 = \frac{328}{337} x100 = 97.329\%$$

The segmentation error of 2,671% is present mostly from the error of splitting a character into two or three character segments.

Then this process is done on all the pages of *Hamong Tani* to obtain the percentage of correct segmentation from each page. Using this information, it is calculated that the average percentage of accuracy of all the pages from *Hamong Tani* is 84.255%, with the standard deviation of 14.093%.

Based on this average percentage of correctness of the segmentation and standard deviation, it can be said that the model developed and used to make the segmentation process on the entire *Hamong Tani* book in this study is acceptable. Results from further investigation shows that segmentation failure may be caused partly by the distortion of the script image due to the scanning process at the start or by the damage of the source page itself.

VII. CONCLUSION

The average percentage of correctness of the character image segmentation system of 84.255% suggests that the method used for segmentation in this study is relatively successful. The model used in this study deserves to be tested on many other Javanese script document images.

ACKNOWLEDGMENT

We are thankful to the Indonesian government for funding strategic research grants nation wide through the Directorates-general of Higher Education in the Ministry of Education and Culture.

REFERENCES

[1] Suryakusuma, S. 2003. "Kamus-Kamus Bahasa Jawa". http://www.tembi.org /perpus/2003_02_perpus03.htm. 23 April 2005.

[2] Srihari, S.N., Lam, S.W., Govindaraju, V., Srihari, R.K., Hull, J.J., 1986, "Document Image Understanding", CEDAR, New York.

[3] O'Gorman, L., and Kasturi, R., 1997, "Executive briefing: documen image analysis", IEEE Computer Society Press., USA.

[4] Kasturi, R., O'Gorman, L., dan Govindaraju, V., 2002, "Document image analysis: A primer", Sadhana, Volume 27: 3-22.

[5] Fletcher, A., dan Kasturi, R., 1988, "A robust algorithm for text string separation from mixed text/graphics images", IEEE Trans. Pattern Anal. Machine Intel., Volume 10: 910-918.

[6] Jain, A.K., dan Bhattacharjee, S.K. 1992. "Text segmentation using Gabor filters for automatic document processing". Machine Vision Appl. J. Volume (5): 169–184.

[7] Wong, K.Y., Casey, R.G., dan Wahl, F.M. 1982. "Document analysis system". IBM J. Res. Dev. Volume (6): 647–656.

[8] Zramdini, Abdelwahab., and Ingold, Rolf., 1993, "Optical font recognition from projection profiles", Electronic Publishing, Volume 6(3): 249-260.