World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:8, 2015

The Implementation of the Javanese Lettered-Manuscript Image Preprocessing Stage Model on the Batak Lettered-Manuscript Image

Anastasia Rita Widiarti, Agus Harjoko, Marsono, Sri Hartati

***Abstract*—**This paper presents the results of a study to test whether the Javanese character manuscript image preprocessing model that have been more widely applied, can also be applied to segment of the Batak characters manuscripts. The treatment process begins by converting the input image into a binary image. After the binary image is cleaned of noise, then the segmentation lines using projection profile is conducted. If unclear histogram projection is found, then the smoothing process before production indexes line segments is conducted. For each line image which has been produced, then the segmentation scripts in the line is applied, with regard of the connectivity between pixels which making up the letters that there is no characters are truncated.

From the results of manuscript preprocessing system prototype testing, it is obtained the information about the system truth percentage value on pieces of Pustaka Batak Podani Ma AjiMamisinon manuscript ranged from 65% to 87.68% with a confidence level of 95%. The value indicates the truth percentage shown the initial processing model in Javanese characters manuscript image can be applied also to the image of the Batak characters manuscript.

***Keywords*—**Connected component, preprocessing manuscript image, projection profiles.

I. Introduction

INDONESIA has thousand manuscripts as the literature asset heritance spread in many area in Indonesia, moreover, they spread abroad [1], [2]. One of efforts which can be done in order to perpetuate the literature asset is by digitalizing all of the manuscripts. By digitalizing the manuscripts, we can get the safer manuscripts saved data and it also makes the manuscripts are possible to be distributed in order to share the great value contain on them to the young generation.

One of the problems which often come in sharing the literature asset contained on the manuscripts is almost of the manuscript are written in certain region's script and certain vernacular which rarely being used. It makes the young generation difficult to comprehend the manuscript.

On the other hand, the presence of computer based technology which has been sophisticated can be used for the research on its relationship with the computation speed and the computerization services. The manuscript transliteration automatically becomes one of the alternatives of the computer utilization for conserving and inheriting the manuscripts. Baried [3] formulate the transliteration as the changing of the character by character type from an alphabet to another alphabet. There are three stages in transliterating the manuscript which should be passed, i.e. preprocessing stage, characters recognition stage, and the interpretation of the characters recognition stage. The first processing stage is a stage to prepare the input manuscript images so they are ready to enter the second stage. The characters recognition stage will determine the changing result of every character to the appropriate character. The interpretation stage is the analyzing of the identification stage. This paper describe the research result from the first stage of Batak character manuscript image transliteration which aims at preparing the manuscript images so they can be processed on the character identification stage by applying the Javanese character manuscript image pre-processing.

A. R. Widiarti is with the Department of Informatics Engineering, Sanata Dharma University, in Yogyakarta, Indonesia (phone: +6281328341628; fax: 01-0274-886529; e-mail: rita_widiarti@yahoo.com).

A. Harjoko, Marsono, S. Hartati are with the GadjahMada University, in Yogyakarta, Indonesia (e-mail: aharjoko@ugm.ac.id, marsono@ugm.ac.id, shartati@ugm.ac.id).

II. The Model of Javanese Character-Manuscript Image Preprocessing Stage

The preprocessing of the manuscript images is started by changing the manuscript image color intensity to the binary images using the Otsu binary method. Binary process is conducted because the color information is not needed in the identification stage. Binary image which has been produced then being treated in order to eliminate the noise which can disturb the next processes. After getting the manuscript images which are relatively free from the noise, the process is continued to the line segmentation. The main approach which is used for line segmentation is by applying the vertical projection on the manuscript images. Since the images are in form of hand writing, the level of the variability of character image lines are high so the vertical projection histogram has the unclear phase space. Reference [4] shows a simple moving average algorithm for produced a histogram that more subtle and clear distance between phases. Histogram segmentation of the new line, and then used as a guide to cut the lines of the image. Each line of the cropped image then is tested whether it has the reasonable height. High-line image is reasonable if its value is under the average height of all objects in the manuscript coupled with a high standard deviation line. On the line image which has a high morbidity of the line, cutting a rough guide is based on the high fairness. To avoid cuts and fix the wrong image, then after cutting the line image, it is conducted the checking operation using the connectivity

between objects connected component. The next step is to cut images of characters in the same line with the projected horizontal histogram instructions. If the images of impropriety segmentation results are high or wide, then the segmentation results need to be processed again. Further processing is conducted to separate the objects that are considered 1 image, when in fact these objects should be divided into two or more images. Operation connectedness becomes a major tool in the further processing to find relationships between pixels in an object.

Broadly speaking, a series of processes to obtain Batak character images is illustrated in Fig. 1.

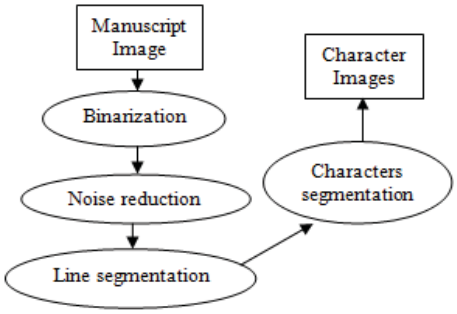

Fig. 1 The series of preprocessing Batak manuscript image

Once the input manuscript image is determined, the first process is conducted the binarization process to obtain a binary image of the manuscript, with the aim of separating the image of the object from the background image. Noise reduction is required to dispose of the objects which are functioned as the background or noise. After the manuscript image free from noise, the next step is the segmentation lines to obtain images of rows of input manuscript image. Results of the line image are then segmented again in characters segmentation that will be produced the character images that make up the manuscript characters concerned.

The whole series of processing in Fig. 1 can already be applied to the image of the Javanese-lettered manuscript, because the characteristics that are common to each line and script in Javanese lettered-manuscripts have a limit boundary line and characters. From the research result of the Batak script writing layout characteristics in the manuscript was shown that there is a similarity between the layouts of the Batak script writing to the Javanese script writing. This study presents the results of research implementing the initial processing of Javanese lettered-manuscript image on the image of Batak lettered-manuscripts.

III. The Analysis of the Batak Lettered-Manuscript Script Segmentation Model Testing

A. Data Sources

The data used to test the Batak lettered-manuscripts segmentation model is the image of the manuscript PodanisiAjiMamisMa Inon, as shown in Fig. 2. The manuscript was downloaded from the blog owned by Rivanca [5], which can represent the Batak Toba lettered-manuscript as information which accompany the manuscript in his blog. For example, test data, research manuscript image in Fig. 2 consists of 12 lines of the image. Because the study only focuses on the character image segmentation, then as the sample used three lines of the input image as shown in Fig. 2 which the section is marked with the white lines are dashed and indicated by arrows.

Fig. 2 PodanisiAjiMamis Ma Inon Manuscript [5]

B. Testing Result

The first step begins with preprocessing binarization process, followed by the reduction of noise in the binary image. In this experiment, not all lines of characters are used, then the process of cutting the image of the manuscript. After obtaining the manuscript binary image that is free from noise, the next step is the segmentation lines with vertical projection, which is a step to get data images making up the image of the input line. The results of the vertical projection of the input image in Fig. 2 are already marked; it is a curve as shown in Fig. 3. The curve in Fig. 3 shows clearly that there are three phases of the curve, so it can be presumed that the input image is clustered into 3 parts.

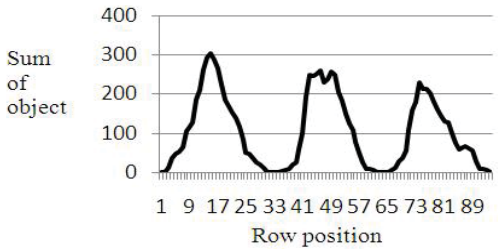

Fig. 3 The curve of the input image vertical projection result

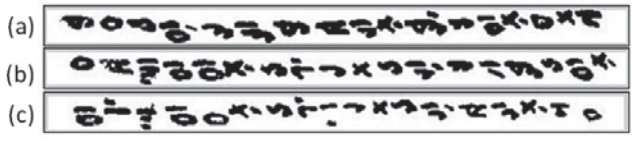

Fig. 4 (a) First row, (b) Second row, and (c) Third row of the result of the image line segmentation

After the cutting process using the information shown in vertical projection result curve is conducted, it is obtained the 3 images as in the Fig. 4.

The process is continued in order to obtain the character images from each line image which has been produced by line segmentation. The first step in characters segmentation is doing horizontal projection of each line image. Fig. 5 displays the yield curve projection of the image of the first row.

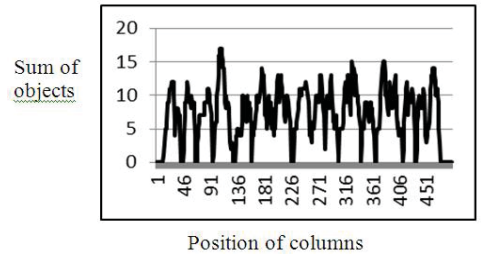

Fig. 5 The curve of the one line image horizontal projection result

Horizontal projection results in Fig. 5 shows that the image in the row has 12 sections which are then considered as the 12-image script, as shown in Fig. 6. An asterisk is located at the top of the cropped image of the script indicates that the result of cutting the image in the wrong section because it contains two or more characters image.

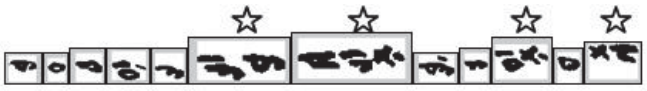

Fig. 6 The curve of the first line image horizontal projection result

Based on the results of the average width of the object calculation which is contained in the input image, in this case amounting 9:29 4:08 pixel with a pixels standard deviation, it can be determined which character images that have unnaturally wide. An image of the characters that have a width exceeding the sum of the average width and standard deviation, it must be certain that the image of the character is a combination of two or more image alphabet. If the image of the original characters in the group unnatural image is not connected, then the operation connectedness images that form a single character can be cut to get one letter image only.

C. Testing Result Analysis

After all tests on the lines of the image is conducted, then in accordance with the ultimate goal of this research is to find any images of character that make up an image of the manuscript thus ready to be processed to the next stage. Based on that goal, the end result of the process in the preprocessing system is produced images of characters forming an image of the manuscript. The research is ended up by determining whether the production of character images at the end of the character segmentation has the true or false value. Character image segmentation results are categorized as the correct one if the letter contains only one image and intact. Table I shows the details of the results of the letter segmentation that have been produced where the input is the line image 1 as shown in Fig. 4 (a).

TABLE I
THE EXAMPLES OF CHARACTERS LIST OF SEGMENTATION RESULT

Character Image	Note
	Correct
	Correct
	Correct
	Correct
	Correct
	Correct
	Correct
	Correct
	Correct
	Wrong because 2 characters are connected
	Correct
	Wrong because the characters are separated
	Wrong because the characters are separated
	Correct
	Wrong because the characters are separated
	Wrong because the characters are separated
	Correct
	Wrong because 2 characters are connected
	Correct
	Correct

Segmentation fault may occur because the original image was already cut or connected, and there is no part of the next character image entered in the current image.

From the automatic segmentation trial on the Batak manuscript segmentation as shown on the Fig. 2 which contains 62 characters, it is achieved the information of which characters are segmented correctly and which segmented wrongly according to the experts. Fifty-four characters are selected randomly in order to analyze its segmentation result, and then the total characters which are segmented correctly and wrongly are calculated. It is obtained that 75.93% characters are segmented correctly. This same process is repeated three times and it is gotten the truth percentage data of 77.78%, 74.07%, and 77.78%. From four true percentage data, it is obtained the segmentation truth percentage average of 73.69%.

With a p-value = 73.69% and n =54 , can be presumed confidence interval average percentage of truth V with a 95% confidence level for the value p by using (1) [6].

$$p - Z_{\alpha/2} x \sqrt{\frac{p(1-p)}{n}} < V < p + Z_{\alpha/2} x \sqrt{\frac{p(1-p)}{n}} \quad (1)$$

IV. CONCLUSION AND FUTURE

The result of the preprocessing system testing for the Batak lettered-manuscript image segmentation produces interval value of truth percentage between 65%-87.68%, so it can be

concluded that the pre-processing model of Javanese characters manuscript can be implemented on the Batak character manuscript. The failure of the preprocessing is caused by the original character image data was wrong, for example, two characters which are connected, or a part of characters image entered another characters image. Effects will arise as a result of such failure is the failure of the characters image recognition, unless in the image recognition script is at once corrected errors. Based on the analysis of the failure of the process, it still needs further research to increase the percentage of truth preprocessing.

ACKNOWLEDGMENT

The authors would like to acknowledge the Indonesian Directorate General of Higher Education (DIKTI) and Fund Management Institution of Education (LPDP) from Ministry of Finance Republic Indonesia for supporting this research.

This research was funded by Indonesia's Directorate General of Higher Education through BPPS scholarship and national strategic grants: Letter of Agreement No. 164/SP2H/PL/DIT.LITABMAS/V/2013, Dated 13 Mei 2013 and No. 078/SP2H/PL/Dit.Litabmas/III/2012, Dated 7 Maret 2012, and by Ministry of Finance Republic IndonesiathroughLPDP grants: Letter of Agreement PRJ-326/LPDP/2013.

REFERENCES

[1] Marsono, *CenthiniTambangraras-AmongragaJilid IV*, Yogyakarta: GadjahMada University Press, 2010.
[2] Susantio, D., *Naskah-naskahKuno Indonesia di Mancanegara*, 2010, http://hurahura.wordpress.com/2010/03/02/naskah-naskah-kuno-indonesia-di-mancanegara/.
[3] Baried, SB., Soeratno, SC., Sawoe, Sutrisno, S., andSyakir, M., *PengantarTeoriFilologi*, Jakarta: PusatPembinaandanPengembangan Ba-hasaDepartemenPendidikandanKebudayaan, 1985.
[4] Efstathiou, C.E., Signal Smoothing Algorithms, No Year, http://www.chem.uoa.gr/applets/appletsmooth/appl_smooth2.html.
[5] Rivanca, *Kumpulan NaskahKuno*, 2010,http://rivanca.wordpress.com/tag/kumpulan-naskah-kuno/.
[6] Wackerly, DD., Mendenhall, W., and Scheaffer, RL., *Mathematical Statistics with Applications 7th Edition*, USA: Thomson Learning, Inc., 2008. http://fvela.files.wordpress.com/2012/01/mathematical_statistics_with_applications1.pdf.

Anastasia Rita Widiarti (M'09–M'11), received her Doctorate's degree in Computer Science from GadjahMada University, Yogyakarta in 2014. Since 2000 she has been teaching at the Department of Informatics Engineering, at Sanata Dharma University in Yogyakarta. Her current research interests include Javanese document image analysis and pattern recognition.

Agus Harjoko received the Ph.D. degree in computer science from of the New Brunswick, Canada, in the field image processing and computer vision. Since 1987 he has been teaching at the GadjahMada University in Yogyakarta.

Marsono is lecturer in Department o[1]f Nusantara Literature Faculty of Cultural Sciences GadjahMada University in Indonesia.

Sri Hartati received the Ph.D. degree in computer science from of the New Brunswick, Canada, in the field Artificial Intelligence. Since 1997 she has been teaching at Computer Science Study Program and Electronic and Instrumentation Study Program, at GadjahMada University in Yogyakarta.