Segmentation of Hanacaraka Characters using Double Projection Profile and Hough Transform

Liliana, Singgih Mardianto Soephomo, Gregorius Satia Budhi, Rudy Adipranata

Informatics Department, Petra Christian University, Indonesia
{lilian, greg, rudya}@petra.ac.id

Abstract. In doing segmentation of Hanacaraka character, Javanese ancient character, one of Indonesian's ethnic ancient character in Java island, the difficulties that occur is the inconsistency of the space between lines, the size of the character and the thickness. Inconsistencies between row spacing and letter size are caused by the letters of the pair, the last vowel and consonant letters in one phoneme. While the thickness is inconsistent due to the writing style of the Hanacaraka itself.
Image Preprocessing needs to be done to get input without skew. To improve skewed text documents, we used Hough transforms to predict the edges of the text area. After that, to segment the line and then continue with segmentation of each character, horizontal projection profile is used and then proceed with vertical.
The result of this segmentation method is good for printed documents. Segmentation process of handwriting documents has difficulty because each row in the document is uneven and very tight between the rows. Those matters cause them overlap. When the line segmented wrongly, the entire character on the line will be not segmented as well. This problem can be eliminate using connectivity test. Before this, it need to segment the line with the overlap area. The character part of below or above the main character can be eliminate because it is not connected to the main character.

Keywords: segmentation, Hanacaraka character, projection profile, Hough Transform, image processing

1 Introduction

A culture basically has a wide variety of variations. In general, the various types of culture are dances, songs, local games; local languages etc. one variation of cultures which is also very visible and often used is the local language. According to a source from Kompas 2012, in continuing study, which was taking samples at 70 sites in Maluku and Papua, the number of languages and sub languages across Indonesia reached 546 languages [1].

Along with the development of technology and global communication, the condition of our culture has been increasingly eroded. Hundreds of local languages in Indonesia are threatened to extinct. There is an estimation of 746 local languages in

Indonesia, yet only more than 400 languages and sub languages has been successfully mapped [2].

Letters in local languages are knows as a form of writing or a representation of that local language. One of the languages having special letters as a form of writing of that language is Javanese with Javanese writing or better known as Javanese characters. The Javanese letters, also known as Hanacaraka and Carakan, is one of the Indonesian traditional characters, used to write Javanese language. In daily lives, the use of Javanese characters is generally replaced with Latin letters which were first introduced by the Dutch in 19th century [3].

Nowadays, there have been enough efforts to preserve the Javanese letters, either by the government or professional circle. One of the good efforts is the development of android-based Hanacaraka application teaching Javanese characters. According to Tekkomdik, besides developing the application, the digitalization of cultural contents such as puppets, macapat songs and also documentary video, would also be launched [4].

2 Theories

2.1 Hough Transform

The data input, in the form of captured image from a digital camera, has the tendency to slant or skew. The skew found is all deviation of the image causing the result after the process of inputting using the hardware differs from the initial image or shape [5]. To solve this problem, Hough Transform is the method used to detect the skew at the image [5] - [8].

Hough Transform is a technique of edge linking and boundary detection, commonly used in image processing [5] - [8]. The purpose of this method is to find the shape of the object in a class of objects using the voting procedure. This voting procedure is conducted in a parameter from the object candidate obtained as local maxima. This parameter will later be called accumulator, specifically formed in the algorithm to calculate Hough Transform.

Fig. 1. is representing the geometric interpretation from parameter θ and ρ. A horizontal line has $\theta = 0°$, with ρ having positive value. A vertical line also has $\theta = 90$, with ρ having positive value at intercept y or $\theta = -90$ with ρ having negative value at intercept y.

The unique calculation concept from Hough Transform is the grouping of parameters ρ and θ into an *accumulator array*. The distance expected at that parameter is $-90 \leq \theta \leq 90$ and $-D \leq \rho \leq D$, where D is the maximum distance between opposite ends in an image. The following steps are to calculate Hough Transform:

1) Perform a Looping for all pixels at the input image. For every non-background pixel P_{ij}.
2) Perform the looping from –D up to D. The mathematical eq.1 to calculate the value of D is as follows:

$$D = \sqrt{((image\_height) \wedge 2 + (image\_width) \wedge 2)} \quad (1)$$

3) Calculate the value of ρ for each angle of $-90 \leq \theta_i \leq 90$.
4) Do the rounding for the value of ρ using the mathematical eq.2.

$$\rho = x \cos(\theta) + y \sin(\theta) \quad (2)$$

5) Do addition at Hough Matrix H_{ij}

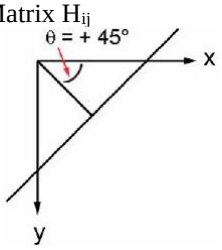

Fig. 1. Representation of a line.

2.2 Segmentation Based on Projection Profile

Projection Profile is a histogram consisting of the number of foreground pixels accumulated along the parallel line in a document [9] - [13]. In several other cases, Projection Profile was used to skew estimation, text line segmentation, page layout segmentation, etc. [9]. The implementation at this application is by dividing the Projection Profile into two types. They are horizontal projection profile and vertical projection profile. The horizontal projection profile is used to find the line region from the document, whereas the vertical projection profile is used to take the character out of each line.

Below is the mathematical equation for horizontal projection profile and vertical projection profile:

$$HPP(y) = \sum_{1 \leq x \leq n} F(x, y) \quad (3)$$

$$VPP(x) = \sum_{1 \leq y \leq m} F(x, y) \quad (4)$$

The samples of the horizontal projection profile and vertical projection profile images can be viewed at **Fig. 2** and **Fig. 3.**

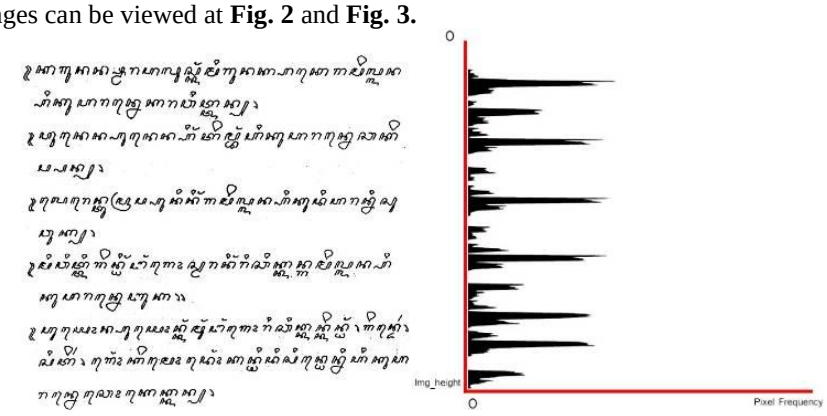

Fig. 2. The Input of digital image and its result of horizontal projection profile

Fig. 3. An image cut of the first row from Fig. 2. and its vertical projection profile

3 Analysis

The main problem in the segmentation of Javanese characters will be resolved in this study is the skew of the documents as the input, the italics and the overlapping writings between lines due to the inconsistency of line spacing or characters sticking together, see **Fig. 4**. The skew document is a problem that often occurs in DCR (Digital Character Recognition) application [5] – [8]. The cause of the problem is the error that occurs during inputting. The skew that occurs is generally $-15° \leq x \leq 15°$.

It has been found that there have been many overlapping or sticking writings condition in a document with Javanese characters. Italics and overlapping writings have brought on the result of the segmentation less optimal. The Overlapping writings in this document were found horizontally (overlap writings between columns) and vertically (overlap in lines). This problem occurs because characters such *as vowel* (special character put above the main character), and *sandhangan* (special character added below the main character) and *carakan* (special adding character for phoneme which adopted from foreign language). Meanwhile the slanted writings are found due to the writing style of the Javanese characters itself. The slanted writing style is not something unusual, as currently some normal texts have slanting writings form which are often called italic. Italics are usually found in a handwritten document.

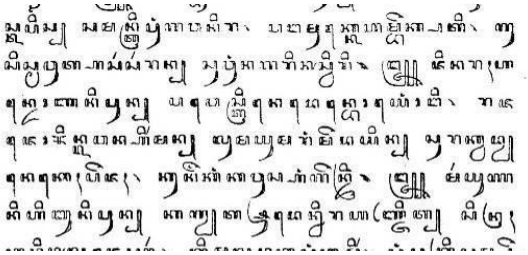

Fig. 4. The sample of an image with overlapping characters in lines

Proposed Method

Fig. 5 shows the system we developed. Some pre-processing need to be performed to get a non-skewed binary image as input for the segmentation process. From "bitmap to array" until "binary thresholding process" are the pre-processing. After get a binary image, the first step is repair a skewed input. Projection profile cannot be performed on a skewed image. To detect and correct the skewed document, Hough

transform is used because it has better performance than scanline method [5]. This method is will determine the border of text area [5] – [8].

If the input has inconsistency space between raw, then the system will run segmentation process without line segmentation. To improve the quality from segmentation without line segment, it will perform filling region procedure. This procedure tries to reconstruct the missing part after the segmentation.

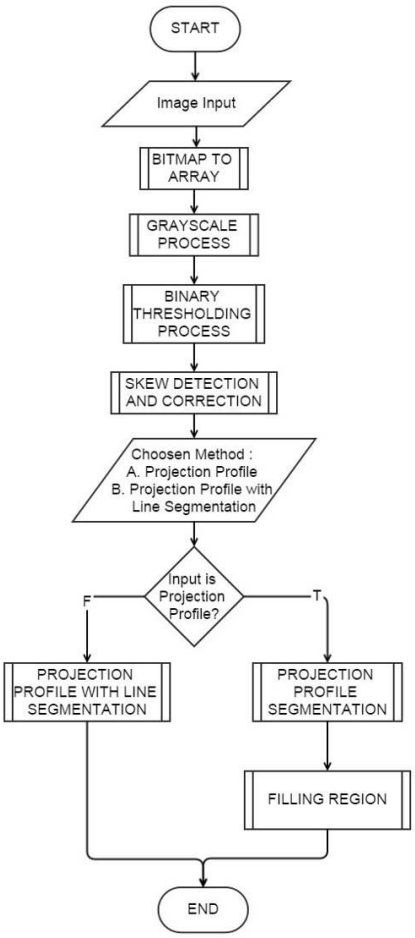

Fig. 5. Flowchart of the system

The other problem is overlapped writings. To solve this problem, there have been several studies conducted. The former study on several kinds of character, such as Kannada [6], [11], Devanagari [7], Arabic [8], Urdu [9], Gurmukhi [10], Chinese character [12] and Oriya [13]. These study ware using projection profile and connected regions methods to do the segmentation of the characters in the writings. However, in several cases, projection profile in the segmentation of Javanese characters cannot be fully applied. The structure of the writings and the unique characteristics of Javanese characters can make projection profile fail.

Kumar repairs the corrupt character with water reservoir technique to get a good result [10], Mamatha uses morphological operation [11], Tripathy uses line segmentation specifically to detect the writing where the line of the writing could not be found using the projection profile, as the writings were overlapped or touched the lines under. Our system uses double projection process to improve the quality of character segmentation.

4 Experiment

The testing performed was to compare the output of the program against the manual calculation. In this testing, the sample of data were classified into two groups, data of the inconsistent spaces between lines (or overlap rows) and data of the inconsistent size and type of characters.

The Inconsistent space between lines

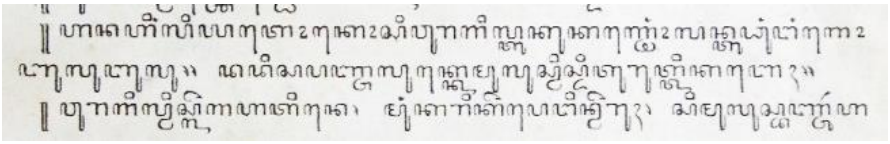

Fig. 6. Photograph of script with line inconsistency [14]

At the sample data as seen at **Fig. 6**, as the first process of projection profile was conducted, only 43 out of 558 characters could be segmented. At the second process of projection profile, 96 out of 558 characters could be segmented.

The Inconsistent Sizes and Types of Characters.

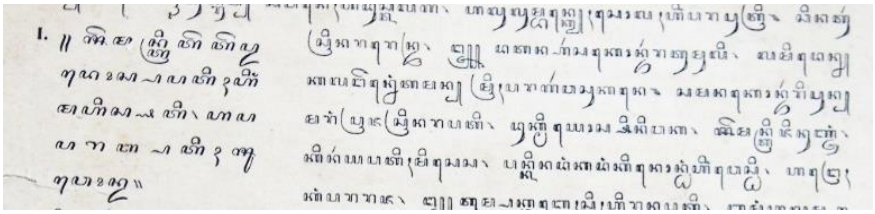

Fig. 7. Photograph of script with size inconsistency [15]

The example of sample data taken from the photograph belonged to the sample data having differences in sizes and types of characters. At the sample data as seen at **Fig. 7**, the writings with different sizes and types could not be segmented as the lines at different parts of the writings were overlapped with other writings. The average result of the testing showed that 15 % of the writings could be segmented. **Table 1** shows the result of segmentation using data in **Fig. 7.**

Using sample data with certain condition, inconsistency space between rows, different font size such as shown in **Fig. 7** and different thickness as shown in **Fig. 8**. Usually, when the hand stroke upper, the line will thinner than when the hand moves lower. Some handwritten documents or even printed documents will have this writing

style. On a printed document, this kind of style will not lead to failed segmentation process because every stroke separated well.

For some document like shown in **Fig. 7**, cannot segmented because some lines of the document laid in differently with the main part.

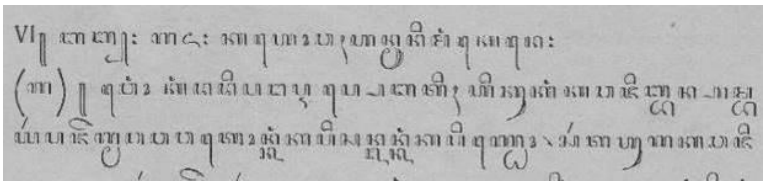

Fig. 8. Hanacaraka writing style [16]

Table 1. Testing using data in **Fig.7**

No	Line on the sample	Writing	Line Output		Writing Output		%
			Right	Wrong	Right	Wrong	
1	25	814	8	17	254	560	31
2	22	558	1	21	79	479	14
3	12	249	12	0	246	3	98
4	14	396	7	3	222	174	56
5	23	805	0	23	0	805	0
6	30	1560	1	29	55	1505	3
7	28	1537	1	27	28	1509	1
8	26	1375	4	22	208	1167	15

5 Conclusion

This system performs better with good hanacaraka image. Printed text images have a constant space between row or not enough space to separate rows, consistent font size and thickness. Those kinds of input will yield good result. The result shown in Table 1 is come from problematic input. Some of them have inconsistency space between rows, and some other have different font size or different thickness in one single character. Overlap rows will lead to fail line segmentation. This will affect the character segmentation also.

Based on the testing result, the projection profile method, on average can perform the segmentation of the writing at a document by 22 % for the group of photograph data having inconsistent spaces between lines. 77 % for the group of photograph data having consistent spaces between lines.

Filling region procedure can help the segmentation to reach 63.5% of character overlap segmentation.

The other problem come from skewed document can be resolve if the skew less than 95°. Document with consistent space between row can be segmented over 75%. Different thickness can be solved as long as each character separately well.

Acknowledgment

This research was funded by DIPA Directorate General of Research and Development Reinforcement (Direktorat Jenderal Penguatan Riset dan Pengembangan) no. SP DIPA-042.06.1.401516/2017, fiscal year 2017.

References

1. Akuntono, I. (2012, September 1). Nasional - Kompas. Retrieved from Kompas Cyber Media: http://nasional.kompas.com/read/2012/09/01/12030360/Mau.Tahu.Jumlah.Ragam.Bahasa.di.Indonesia
2. Sudirman, M. (2013, Maret 5). Republika. Retrieved from Republika: http://www.republika.co.id/berita/koran/news-update/14/03/04/n1wzn0-bahasa-daerah-semakin-punah
3. Agfa Monotype Corporation. (2000, January 1). Monotype. Retrieved from Monotype: http://www.monotype.co.uk/NonLatin/wt_info/info_javanese.html
4. Nbi. (2013, Desember 20). Hanacaraka, Aplikasi Android untuk Belajar Aksara Jawa. Retrieved from Tribunne News Cyber Media: http://jogja.tribunnews.com/2013/12/20/hanacaraka-aplikasi-android-untuk-belajar-aksara-jawa/
5. Kishan, A. C., & Sharda, V. (2009). Skew Detection & Correction in Scanned Document Images. Orissa: Department of Computer Science and Engineering, National Institute of Technology Rourkela.
6. Ramappa, M. H., & Srikantamurthy, K. (2012). Skew Detection, Correction and Segmentation of Handwritten Kannada Document. International Journal of Advance Science and Technology, 71.
7. Rahul Garg, Naresh Kumar Garg. (2014). An algorithm for text Line Segmentation in Handwritten skewed and Overlapped Devanagari Script. International Journal of Emerging Technology and Advanced Engineering 4(5) : 114-118
8. Yasser M. Alginahi. (2013). A survey on Arabic character segmentation. International Journal on Document Analysis and Recognition. 16(2):105–126.
9. Javed, M., Naghabushan, P., & Chaudhuri, B. (2014). Extraction of Projection Profile, Run-Histogram and Entropy Feature Straigt from Run-Length Compressed Text Documents. Kolkata: Department of Studies in Computer Science, University of Mysore.
10. Kumar, M., Jindal, M.K., Sharma, R.K. (2014). Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition. I.J. Information Technology and Computer Science, 2 : 58-63
11. Mamatha, H.R, Srikantamurthy, K.. (2012). Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document. International Journal of Applied Information Systems (IJAIS) 4(5) : 13-19
12. Mei, Y., Wang, X., & Wang, J. (2013). A Chinese Character Segmentation Algorithm for Complicated Printed Documents. International Journal of Signal Processing, Image Processing, and Pattern Recognition, 6 (3):91-100
13. Tripathy, N., & Pal, U. (2006). Handwritting Segmentation of Unconstrained Oriya Text. Sadhana, pp.755-769.
14. BPAD Tentara Pelajar. (1992). Dongeng Koetjing Setiwelan. Yogyakarta, p. 5
15. BPAD Tentara Pelajar (1938). Langendriya, Yogyakarta, p. 7
16. Rijksblad (1936), Yogyakarta No 1. p. 9.