International Journal of Engineering Innovation & Research
Volume 2, Issue 3, ISSN: 2277 – 5668

Line Segmentation of Javanese Image of Manuscripts in Javanese Scripts

Anastasia Rita Widiarti
Email: rita_widiarti@usd.ac.id

Agus Harjoko
Email: aharjoko@ugm.ac.id

Marsono
Email: marsono@ugm.ac.id

Sri Hartati
Email: shartati@ugm.ac.id

***Abstract* – Segmentation is an important stage in automatic transliteration process of a manuscript image. One of the segmentation approaches generally used to get an image of the scripts of a scrip image is performing line segmentation and then performing segmentation of script images on the result of the line segmentation.**

Line segmentation of image of manuscripts in Javanese scripts is often difficult because there are lines of images which shouldn't be on the same line, but are in one line area of the image, and there are even images in different but overlapping lines. This paper offers an idea to use moving average to smooth the curve of vertical projection result of image of manuscripts in Javanese scripts as an initial guide of line separation. Next, to separate parts of scripts in different lines but in the same line or in overlapping lines, average height data and standard deviation of average height of every objects are use to get information on estimation of the height of line image. Image connectivity concept in an object is also used to separate script images in different but overlapping lines.

From the test result of 4 images of manuscripts in Javanese scripts from different writers with different writing styles, average percentage of correctness of line segmentation obtained was 93.19% with standard deviation of 4.55%. Mistakes of line segmentation were mostly caused by *sandangan* of a script in the previous line which joined the next line, and cutting imperfect scripts. However, from the percentage of correctness of line segmentation, it could be concluded that the combination of various image processing techniques used in line segmentation was relatively good.

***Keywords* – Connected Component, Javanese Manuscript Segmentation, Moving Average, Projection Profile.**

I. INTRODUCTION

Segmentation of script image has a very significant and major role in the efforts to introduce a document image. The success of a segmentation process will influence the next processes, especially script introduction process. Moreover if the manuscript which will be recognized is a hand-written manuscript. This is a challenge, because manuscripts are written by many different writers, and the writing style may have different *gagrak*.

Segmentation of script image can be done by first performing line segmentation of manuscript image, and then after obtaining lines of manuscript image, it's continued by segmentation of script images from corresponding line of manuscript image. However, line segmentation of manuscript image in often difficult because there are fluctuation lines, overlapping different components, and irregularity in geometry of the lines, such as the height and width of the lines [1].

II. RELEVANT WORKS

Palakollu, et al. [2] investigate line segmentation method on Hindi manuscript image using projection-based approach. They build an algorithm to detect header line and base line, based on several initial assumptions such as average line height of 30 pixel, to make an estimation of real average line height. From 500 documents investigated, the average accuracy of line segmentation is 93.6%. Lehal and Singh [3] study segmentation on Gurmukhi text using a combination of statistical analysis from the text, projection profile, and analysis on connected components. By using horizontal projection, they discover segmentation failures which are separated into two failures, which are over segmentation and under segmentation. Over segmentation happens because there is a gap between lines of text, and under segmentation happens because there are parts of a text which overlap between lines. Lehal and Singh then determines the value of certain unit based in characteristics of placement of scripts, i.e. height of an area, determination of strips classes, which are script groups in a line, and determination of the height of the first line. From the test on 40 documents of Gurmukhi printed text, they produce average percentage of correctness of overlapping script segmentation of over 87%. Tripathy and Pal [4] provide a sample of main principle for line segmentation on Oriya manuscript, which is dividing the scrip into several groups vertically the applying horizontal projection on the parts. Relation between the peaks and slopes on histograms are information for performing line segmentation. Structural and topological approaches and characteristics of the curves formed into some kind of reservoir of script image histogram used to detect isolated and overlapping scripts. With the approaches, success percentage in separating overlapping scripts was 96.7%. Nicolaou and Gatos [1] cut lines of manuscripts based on the assumption that there is a connectivity line in a line of manuscript. Weliwitage, et al. [5] uses projection modification named cut text minimization (CMT) to perform line segmentation of constant slope characteristics. The main principle of CMT is finding a line or cutting line by cutting words not on the line as minimally as possible. Yin and Liu [6] perform line segmentation of image of manuscripts in Chinese scripts, based on grouping of connected components using minimal spanning tree (MST), and then performing distance matrix-based learning. During the test of 803 manuscripts in Chinese characters with a total of 8169 lines of line image, the algorithm they proposed has very high percentage of correctness in detecting line, which is 98.02%. Surinta, and Chamchong [7] successfully perform

segmentation of historic image handwritten on palm leaves. It start with binarization stage with Otsu method to separate object and background, then continued with line segmentation stage by applying profile projection method, and finally segmentation stage to get scripts using histogram of segmented images. They get an average of percentage of correctness of scripts of 82.5%. Jindal et al. [8] on horizontal segmentation for lines of overlapping printed Indian scripts successfully get accuracy lever of segmentation of 96.45% to 99.79%.

Widiarti [9] has studied the use of profile projection to get images of Javanese scripts from printed manuscripts in Javanese scripts, with success rate of segmentation of 86.78%. Profile projection in this case is very possible to be used because the characteristic of printed manuscripts in Javanese scripts is having clear and even uniform distance between lines and scripts. In this paper, we have proposed new strategies to line segmentation from Javanese manuscript. We use of vertical projection for line image segmentation on manuscripts in Javanese scripts combined with popular technique to smooth curves, which was moving average method, and combined again with statistical information from data on the height of objects in an image and using pixel connectivity concept on the same object.

III. 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. 1 shows 20 *legena* Javanese scripts.

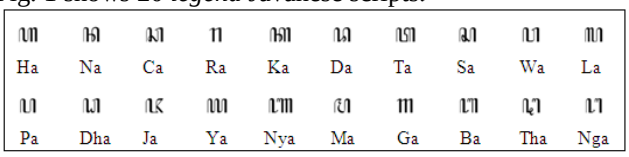

Fig.1. *Nglegena* Javanese scripts [9]

Generally, most Javanese scripts used don't only consist of *legena* scripts, bust 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. 2.

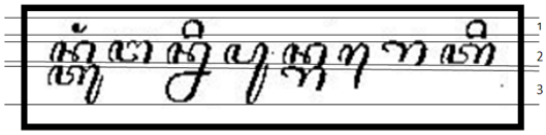

Fig.2. Writing area of Javanese script

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. 3. shows examples of Javanese scripts and places to put the scripts.

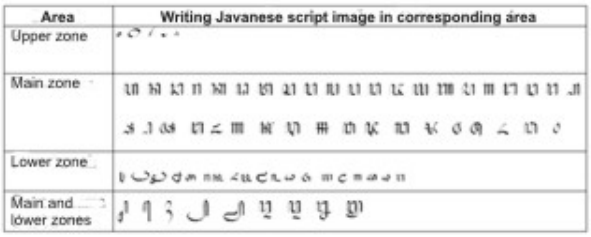

Fig.3. Writing zones of Javanese scripts

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

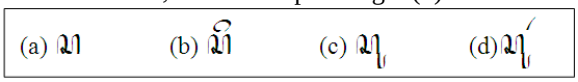

Fig.4. Samples of script forms with combination of the placement of the additions.

From results of a study on methods of writing Javanese scripts on manuscripts, problems related of line segmentation of manuscripts in Javanese scripts are discovered as shown in Fig. 5, because:

1) There are scripts on different lines which overlap. Fig. 5(a). Shows that part of script image in the upper line touch part of script image in the second line. Dashed line is the sign for line change.
2) Distance between scripts isn't clear, so it's impossible to only use vertical projection. Fig. 5(b). shows that part of script image in the upper line is in the same position as script image in the next line.

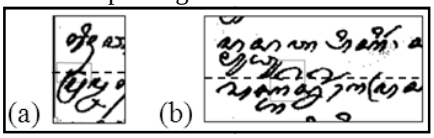

Fig.5. Samples of problems in line segmentation of manuscripts in Javanese scripts.

IV. Solution Proposal

Manuscript segmentation began with finding lines of scripts which formed the manuscript, then finding scripts which form script lines. Once a binary script image was free of noise, then line segmentation was determined.

General working method to get script image segmentation from manuscripts in Javanese scripts with the limitation of characteristics of Javanese scripts writing method is shown in Fig. 6.

To perform crude line segmentation, the tool used first was vertical projection. Because there were manuscript lines which overlap, the result of vertical projection should be refined to get a curve which reflected clarity of distance between phases. Phases in the curve gave location clue of the beginning and end of a line, because 1 phase showed 1 line of manuscript image.

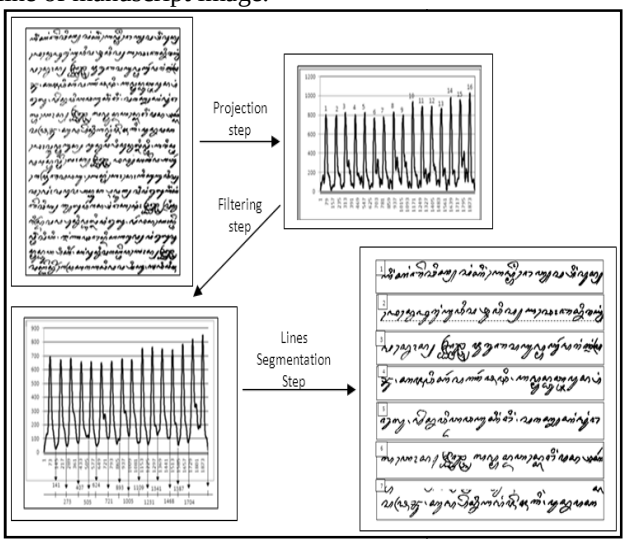

Fig.6. Framework of line segmentation stages

Moving average is a concept often used in statistics to refine curves. If a data set is discovered $[y_1, y_2, \ldots, y_N]$ then a new data set resulted from the refining of the data can be discovered. Equation (1) is one of the formulas to refine one data element (y_k), into (y_k)s [11].

$$(y_k)_s = \sum_{k=-n}^{k=n} y_{k+1}/(2n+1) \quad (1)$$

Equation (1) was adopted in this study to refine curves.

Due to the characteristics of manuscripts in Javanese scripts which often had no clear distance between the lines, the result of line image cutting with vertical projection which had been refined still cut several scripts. For every script which was really in the line range found, the script was clearly in its line, but often there were parts of scripts in the line below them. To solve this, connectivity concept between the pixels was applied so that parts of the scripts could be discovered.

Abnormality appeared when the height of a part of a script, or hereinafter very big object, which was over the average height of standard scripts plus twice standard deviation value of average height. If there was any abnormality in the object height, the object must be two different objects from different lines which were connected, as shown in Fig. 5(a). In principle, the object might be part of scripts in different lines, so it had to be cut or separated.

To calculate the height of every object which might make a script, the distance between the highest position to the lowest position of every object was calculated. After the height of every object was discovered, average object height and standard deviation of average height in the manuscript could be calculated.

The main principle in cutting two different scripts and in different line positions but were connected was by cutting the scripts right at the position which is the distance of average height plus twice the standard deviation of the average height calculated from the beginning of the line. This convention was used because generally Javanese scripts have *legena* scripts and *sandangan* above and below them, as seen in Fig. 4(d). Average object height was seen as average height of *Legena* scripts, while the height of *sandangan* was considered absolute value of standard deviation of average height.

To solve the problem of having parts of a script in the next line which was usually *sandangan* in the upper zone for a script right below it which was in the present line position, the first was is looking for the highest position of the *sandangan*. If the highest position found had a distance above average object height, the object must be *sandangan* in the upper zone of a script in the next line. This convention was based on the belief that *sandangan* of a script in the lower zone didn't have a distance above the average height of the main script. Therefore, the object couldn't be *sandangan* in the lower zone of the script in the line.

V. Result and Discussions

The study started by selecting manuscript which would become testing data, with a contention that the manuscript should be written by different people and with different writing styles from each person. Table 1 shows description of data source and information of data chosen for test data in this study, and Fig.7. shows sample image of a page of script in number 3 in Table 1.

Table 1: description of test data

No.	Catalogue Number / Book Information	Manuscript Storage
1	W74 Pakem Ringgit Purwo (5 lampahan) PB C 69 8 Javanese Language Javanese Script Roll 40 no. 6 [12]	Sonobudoyo Museum Yogyakarta
2	S160 Serat Babad Jumeneng Sultan Kabanaran SK 124 1036 Javanese Language Javanese Script Rol 111 no. 1 [12]	
3	PB A.57 175 Javanese Language Javanese Script Macapat Rol 153 no. 9 [12]	
4	Serat Pertanda	Pustaka Artati of Sanata Dharma University

From the characteristics of manuscript image in Fig.7., it was discovered that there was a gap between lines of the imaged, although in several lines of the image there were

scripts which overlap. By using the information on the characteristics, the first step was vertically projecting data of the image of the manuscript with the scripts. From the data of the result of vertical project, a cover as seen in Fig. 8 was made.

The peaks of the 16 curves in Fig. 8 produced information on the number of line images in the corresponding data, but this curve also showed that the boundaries between lines weren't clear due to high value variation in the slopes of the curves. The vagueness of the distance between lines created a problem in line segmentation, because it made determining cutting position difficult. One of the methods to get a clear gap between lines was by refining curves with moving average. The main principle of moving average method in curve refining is by remapping data of curve values using values around the value which would be remapped. By using this, it was expected that varying data at close range became more uniform or like other data around it.

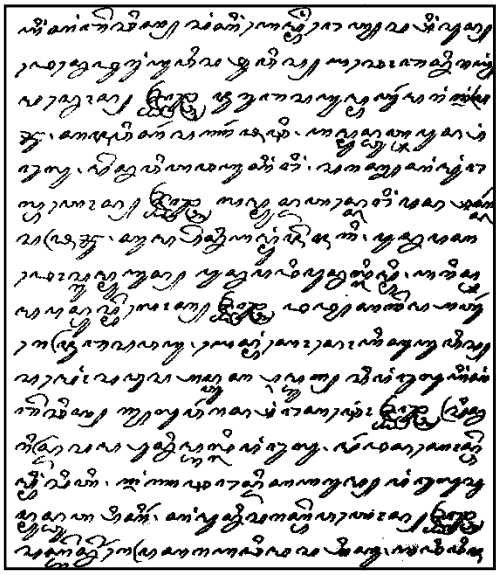

Fig. 7. A section of an image of a manuscript in Javanese scripts, collection of Sonobudoyo museum with catalogue number PB A.57 175 Javanese Language Javanese Script Macapat Rol 153 no. 9 [12]

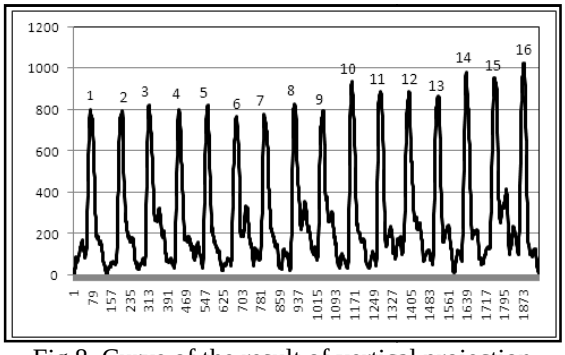

Fig.8. Curve of the result of vertical projection

Fig.9. shows a curve which was refined with slot size 5, it means that the present value is the total of two values at the right and left of the spot and the spot itself, and the refining was performed 6 (six) times.

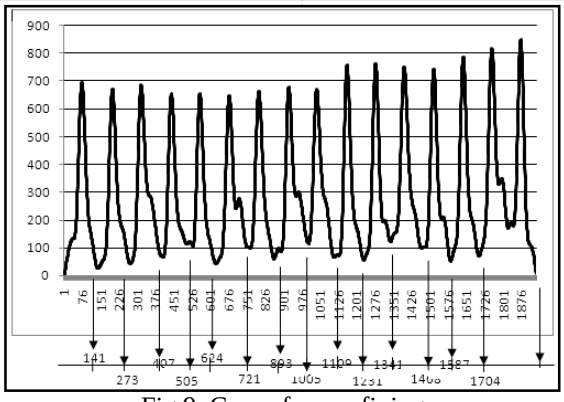

Fig.9. Curve from refining

The slope of the result of the curve refining often didn't refer to the number 0 (zero) which meant that there was a gap between the lines, because in the manuscript there were parts of scripts on the upper line or lower line in the same range, or overlap, or connect. One way to solve this was determining which line had significant data shifting, and this line would be the reference for separating the lines. In this study, the method to determining the line shift is by marking the line in which there was shifting of the result of subtraction of initial value with new value which had been refined from negative to positive or vice versa. The result of the test on line shifting information obtained became the initial clue that there were line changes in the numbers.

Besides that, from the study on the characteristics of the height of every object in manuscript image, statistical information of average object height connected and its standard deviation were determined. Table 2 shows the summary of study result on average object height and its standard deviation on 4 (four) images of manuscripts in Javanese scripts used.

Table 2: Statistical data of object height on test image

No data	Total Object	Average Object Height	Standard Deviation of Average Object Height
1	204	62.505	43.573
2	308	36.325	24.402
3	442	44.267	20.684
4	148	28.905	19.206

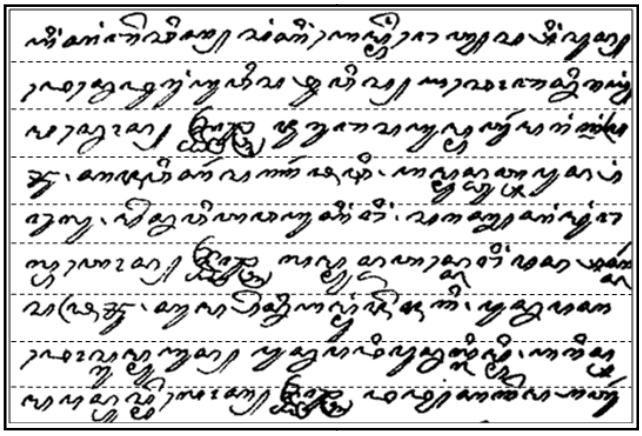

Fig.10. A section of manuscript image with line guide

Fig.10 shows a section of scrip image in Fig. 7 and Fig. 11 shows a sample of segmentation result of images of lines from manuscript image. In Fig. 10 there are dashed lines to show the limit of each line. In the figure it can be seen that starting from the image of the second line and so on, there are parts of script images in different lines but in the same area or overlap.

Fig.11.1. to Fig.11.9. show segmentation result from the models offered, besides Fig. 11.2a. Fig.11.1. shows that segmentation result in line 1 is very clean, it means that the image was segmented well without connection with other lines. This was because there was a clear line in the original image between the first line and the second line. Line segmentation in the next lines, i.e. the second line shows data overlapping. Fig. 11.2a. with dashed lines show that there are parts of the scripts in the third line which enter the area of the second line. By using information on average object height as well as its standard deviation, line separation or cutting could be done.

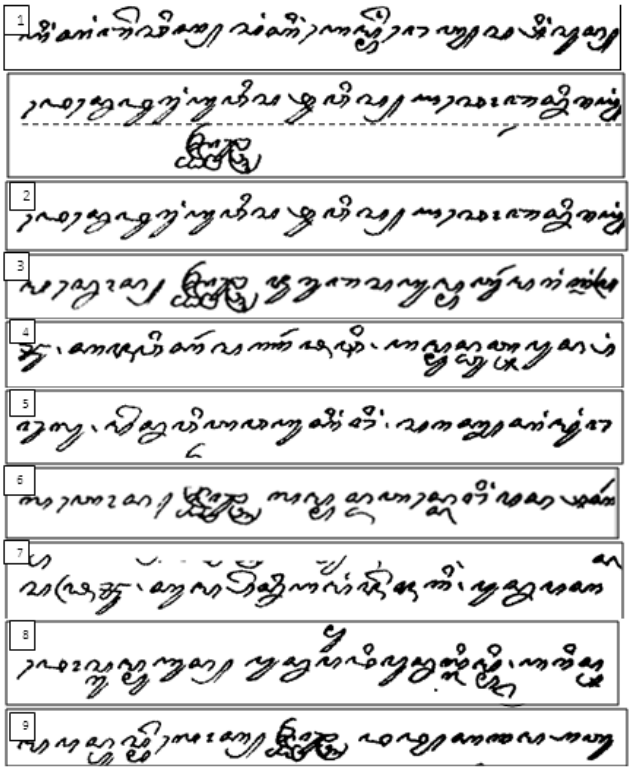

Fig.11. The result of segmentation of the image in Fig. 10

Using visual observation on each segmented line, the number of scripts which should be in a line and still in the line could be calculated. Table 3 shows information of the number of lines in the original image and segmentation result which had been performed as seen in Fig. 11.

Using simple method of calculating percentage of correctness, percentage of correctness of line segmentation for manuscript image in Fig.11 was discovered as follows:

$$\%\ of\ correctness = \frac{Total\ output\ script}{Total\ actual\ script} x\,100 = \frac{151}{159} x\,100 = 94.97 \quad (2)$$

The same experiment was applied for three other images of manuscripts in Javanese scripts. After the result of line image segmentation on the four manuscript images used as testing material was obtained, percentage of correctness of line image segmentation on all test data could be calculated. Table 4 shows the summary of the data of the result of line image segmentation determination in all manuscripts used as test data.

The calculation of percentage of correctness on four different manuscript images used in this study results in average percentage of correctness of 93.19% with standard deviation of 4.55% as seen in Table 4. With very high average value of percentage of correctness, which is more than 90%, it was concluded that the model suggested for line segmentation on images of manuscripts in Javanese scripts in this study was good.

Table 3: Observation data on the number of original scripts and output of segmentation from Figure 11.

Line No	Number of Script		Information
	Original	Output	
1	17	17	All scripts which should be in certain lines are in place
2	20	20	
3	16	16	
4	18	18	
5	18	18	
6	16	11	There were 3 sandangan which joined the precious line, and 2 scripts which were cut imperfectly
7	19	18	There was 1 which joined the previous line
8	19	19	All scripts which should be in the line are in place
9	16	14	There were 2 *sandangan* which joined the previous line
Total scripts	159	151	

Table 4: Summary of percentage of correctness of line segmentation of the four manuscript images

No Data	Percentage of Correctness
1	87.36
2	98.10
3	92.31
4	94.97
Average percentage of correctness	93.19
Average standard deviation	4.55

VI. Conclusion and Future Work

From the experimental test of segmentation model offered, it could be concluded that line segmentation of images of manuscripts in Javanese scripts could be done using guides of vertical projection of the manuscript image combined with moving average refining method with average object height value as well as its standard deviation. Segmentation result could contain several overlapping images, but by using label information of every connected object, images of lines in Javanese scripts connected to it could be found. However, this study could be continued to finish the problems of separating or cutting overlapping scripts, and whether the result of the line segmentation could be used well in script segmentation later.

References

[1] A. Nicolaou, and B. Gatos. (2009). Handwritten Text Line Segmentation by Shredding Text into its Lines. *10th International Conference on Document Analysis and Recognition, IEEE*. pp. 626-630. Available: http://www. cvc.uab.es/ icdar2009/papers/3725a626.pdf.

[2] S., Palakollu, R. Dhir, and R. Rani. (2011). A New Technique for Line Segmentation of Handwritten Hindi Text. *Special Issue of International Journal of Computer Applications (0975 – 8887) on Electronics, Information and Communication Engineering – ICEICE.* Available: http://research. ijcaonline.org/iceice/ number5/iceice033.pdf.

[3] G.S. Lehal, and C. Singh. (No Year). *A Technique for Segmentation of Gurmukhi Text*. Available: http://advanced centrepunjabi.org/pdf/A%20technique%20for%20segmentation %20of%20gurmukhi%20text.pdf.

[4] N. Tripathy and U. Pal. (2006, Dec.). Handwriting segmentation of unconstrained Oriya text. *Sadhana*. Volume(31), 755–769. Available: http://www.ias.ac.in/ sadhana/ Pdf2006Dec/755.pdf.

[5] C. Weliwitage, A.L. Harvey, and A.B. Jennings. (2005). Handwritten Document Offline Text Line Segmentation. *Proceedings of the the Digital Imaging Computing: Techniques and Applications (DICTA 2005)*. Available: http://nguyendangbinh.org/Proceedings/DICTA/2005/data/27_c weliwitage_textsegment.pdf.

[6] F. Yin, and C. Liu. (2009). Handwritten Chinese text line segmentation by clustering with distance metric learning. *Pattern Recognition*. Volume(42), 3146-3157. Available: http://nlpr-web.ia.ac.cn/2009papers/gjkw/gk13.pdf.

[7] O. Surinta, and R. Chamchong. (2008) *Image Segmentation of Historical Handwriting from Palm Leaf Manuscripts*. Available: http://www.wbi.msu.ac.th/file/721 /doc_57.pdf.

[8] M.K. Jindal, R.K. Sharma, and G.S. Lehal. (2007). Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts. *International Journal of Computational Intelligence Research*. Volume (3), 277-286. Available: http://www.ijcir.info.

[9] A.R. Widiarti, "Segmentasi Citra Dokumen Teks Sastra Jawa Modern Mempergunakan Profil Proyeksi," *SIGMA Jurnal Sains dan Teknologi*, vol. 10, 2007, pp. 167-176.

[10] Anonim. (No Year). Available: http://id.wikipedia .org/ wiki/Aksara_JawaW.

[11] C.E. Efstathiou (No Year). *Signal Smoothing Algorithms*. Available: http://www.chem.uoa.gr/applets/appletsmooth/ appl_smooth2.html.

[12] T.E. Behrend, *Katalog Induk Naskah-naskah Nusantara Jilid I Museum Sonobudaya Yogyakarta.* Jakarta: Djambatan, 1990.

Author's Profile

Anastasia Rita Widiarti
received Master's degree in Computer Science from Gadjah Mada University, Yogyakarta in 2006. Since 2000 she has been teaching at the Department of Informatics Engineering, at Sanata Dharma University in Yogyakarta. Her current research interests include Javanese manuscripts image analysis and pattern recognition.

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

Prof. Dr. Marsono
is lecturer in Department of Nusantara Literature Faculty of Cultural Sciences Gadjah Mada University in Indonesia.

Sri Hartati
received the Ph.D. degree in computer science from the of 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 Gadjah Mada University in Yogyakarta.