A proposed model for Javanese manuscript images transliteration

IOP Conf. Series: Journal of Physics: Conf. Series **1098** (2018) 012014 doi:10.1088/1742-6596/1098/1/012014

A proposed model for Javanese manuscript images transliteration

A R Widiarti[1,a], R Pulungan[2,b], A Harjoko[2,c], Marsono[3,d], S Hartati[2,e]

[1]Department of Informatics Engineering, Sanata Dharma University, Mrican Tromol Pos 29, Yogyakarta, Indonesia
[2]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia
[3]Faculty of Cultural Sciences, Universitas Gadjah Mada, Jalan Sosio Humaniora, Bulaksumur, Yogyakarta Indonesia

E-mail: [a]rita_widiarti@usd.ac.id, [b]pulungan@ugm.ac.id, [c]aharjoko@ugm.ac.id, [d]marsono@ugm.ac.id, [e]shartati@ugm.ac.id

Abstract. Manuscript transliteration is generally conducted by reading the manuscript and then writing the results to another piece of paper or storing them on a computer using a specific text-processing program. The procedure of transliteration presupposes that the workers fully understand how to read the manuscript, that no consideration is given to the length of the work, and that workers will concentrate sufficiently to minimize errors in rewriting. There are 3 main steps for Javanese manuscript image transliteration, *i.e.*, segmenting manuscript, transliterating of Javanese script letters or numbers, and grouping syllables. The implementation of the proposed model was tested on a Javanese manuscript with catalogue number SB.141 and at a confidence level of 95% and resulting in a success rate between 69.20% and 87.29%.

1. Introduction

As a great nation, Indonesia has a long cultural heritage captured in the form of information about the noble culture of the past, in the form of either scripts or manuscripts that remained uncounted until recently and exist in various locations [1]. These manuscripts contain historical information about worship, manners, history, folklore, traditional technologies, spells, genealogy, talismans, poetry, politics, governments, laws, customs, traditional medicine, and culture [2]. The manuscript media take various forms. The manuscripts were written in many different languages. Moreover, many different scripts were used to write the manuscripts [3, 4, 5, 6].

Transliteration is the replacement of writing, letter by letter, from one alphabet to another [7]. The purpose of transliteration is to make old texts written in local script more accessible, because most people are no longer familiar with the original script, such as old Javanese scripts. However, transliteration usually cannot be conducted quickly except by philologists because of the difficulties involved in reading the local script. This paper describes the results of research on automatic transliteration, which is an implementation of the transliteration model for Javanese manuscript images using a specific programming language as an alternative method to solve the problems involved in the transliteration.

Many studies have been conducted that are related to transliteration either in whole or in part. Shridhar and Kimura [8] provided the steps to identify handwritten words in Roman

script. The word recognition process begins with correcting any slant in the document's image and then analysing images of the rows of words in the document. After obtaining the row images, the process continues by first correcting any slant in the line and then acquiring images of the individual words in those rows. An initial segmentation process is conducted for every obtained word image to obtain the basic components that form the words. The final step is to incorporate the basic components obtained in the previous processes by recognizing each character forming the words. Tangwongsan and Sumetphong [9] developed a character recognition system to analyse images of historic documents from Thailand, which achieved a success rate above 90%. There are three main stages in the developed system: data preparation, segmentation and character recognition. The data preparation stage includes two main processes: document image binarization using the Otsu method, and fixing any slant in the document image using the Hough transformation method. During segmentation, two methods are applied: profile projection and object cutting. During the recognition stage, Daubechies wavelets method is applied and followed by a principle component analysis.

The document image recognition methods described in [8, 9] imply that the stages in document image recognition methods recognize the contents of the document image in letter-by-letter fashion. Based on the studies above the manuscript transliteration approach used in this study also recognizes Javanese manuscript image in letter-by-letter fashion.

2. Challenges

The first challenge of manuscript image preprocessing is that manuscripts to be digitized are often dirty and the digitization is incomplete. The manuscripts that must undergo digitization are generally fragile due to the paper used to write the original manuscript or imperfect storage conditions, which result in a non-white or yellowish colour, spots and consequently, unclear script images. Another challenge stems from the characteristics of the script image of the tested manuscripts itself: the boundaries between script rows are not clear, causing the script to appear to contain only 1 object rather than a group of 2 or 3 objects.

The second challenge is that the handwriting found in manuscripts tends to be inconsistent in terms of thickness, size, and the slant of the script. This is likely due to differences in the pens used, the pressure applied when writing, the style used to grasp the pen, and the use of inks that may have had different drop volumes.

Finally, Javanese manuscript transliteration requires that the process is not be limited only to script replacement but also includes the formation of the words resulting from the script replacement. Here, a new problem arises, because classic Javanese scripts did not use space symbols to separate words.

3. Proposed Method

After conducting a survey to identify the data sources and performing a literature search related to the concepts of automatic transliteration and the rules of Javanese script writing, a model of transliteration is created, followed by the establishment of a Javanese scripts feature and its text label that is used to test the model implementation in the laboratory. During testing, whenever the results achieves a very low success rate, the model and the pertinent modules are revised and retested, until the transliteration process achieves a fairly good success rate. This paper extends the result of [10], especially in the postprocessing phase when manuscript images are being recognized.

Figure 1 depicts the details of the general transliteration model along with detailed components of each subprocess or submodel. There are three main parts in the complete manuscript transliteration system: (A) establishing a database of words, syllables, and a probability table, (B) establishing a database of script characteristics, and (C) image transliteration of the manuscript.

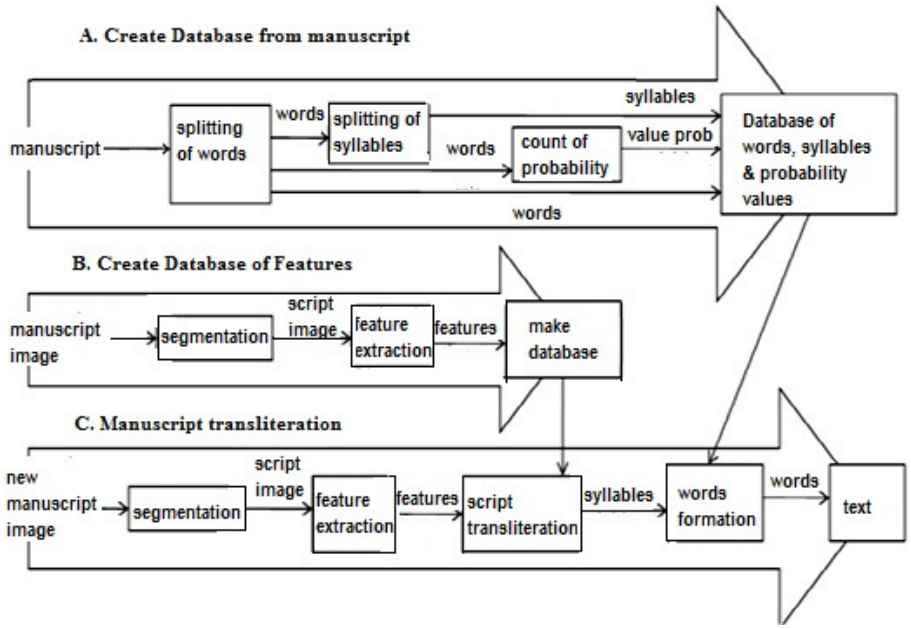

Figure 1. The general model of manuscript image transliteration.

The goal of Section A in Figure 1 is to produce a database of words and a probability table showing the probability of the appearance of a syllable after another syllable or a word after another word. The input of this section is the Latin manuscripts. The manuscripts are in .txt format and form rows of sentences that will be fragmented into rows of words during the "Word Fragmentation" subprocess. Then, those words will be separated into syllables during the "Syllable Fragmentation" subprocess. After producing a row of words and syllables, the probability tables showing the likelihood of the appearance of a syllable after another syllable are produced. Similarly, the probability tables showing the likelihood of the appearance of a word after another word are created in the "Probability Calculation" subprocess.

The goal of Section B in Figure 1 is to produce a Javanese script image database, which contains data representations of the images of Javanese script, as well as the features extracted from each Javanese script image as well as descriptions of the names of these characteristics in the form of syllables. The inputs for section B are the manuscript images that will be used as the main materials when conducting transliteration of the input Javanese script images.

Section C is the most important part in the overall transliteration model of manuscript images. The input of this section is the manuscript images that will be transliterated. Three subprocesses must be conducted to obtain the transliteration of the input manuscript image: segmentation, script image transliteration, and word formation. The expected output of the segmentation process is the rows of script images that form the input of the manuscript images. The Javanese script images are transliterated into Latin script during script image transliteration. Finally, the transliterated script image will undergo the word formation process to obtain a row of words in text format.

3.1. Manuscript segmentation

The goal of manuscript segmentation is to get all Javanese script images from the input manuscript image transliteration. The input for processing consists of the scanned images of the copy of the original manuscript. Based on characteristics of the Javanese manuscript image—*i.e.*, each of row is separated by space, and each of script is separated by space too—

then projection profiles are used for the segmentation of the manuscript image. The complete processing model for script segmentation is depicted in Figure 2.

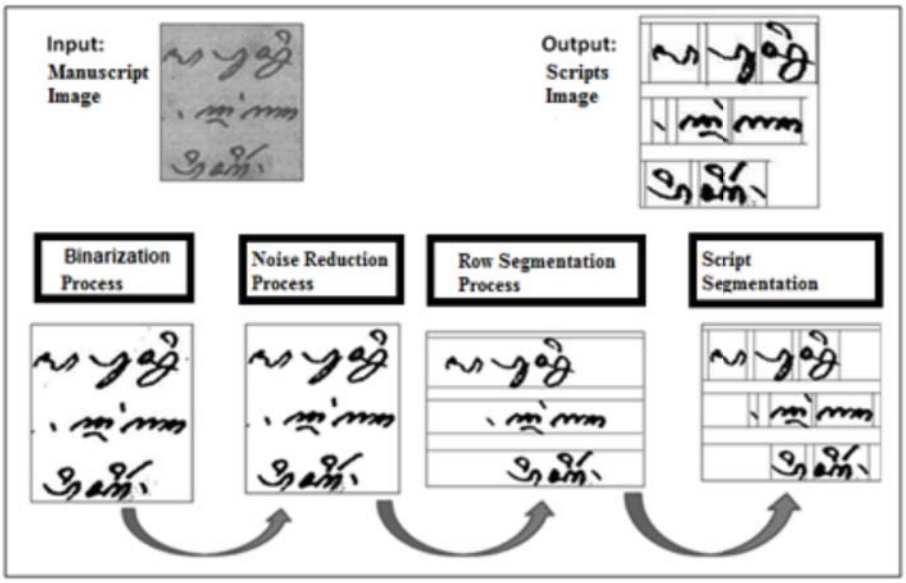

Figure 2. The model for segmentation of Javanese manuscript.

First, the image is binarized, which converts the input image into a binary format containing only black or white pixels. The Otsu method was selected for binarization because it produces good results on printed text document images in Javanese script.

Observation of the new binarized image shows that considerable noise exists, which looks like groups of dots or grains of sand. Therefore, the next step is noise reduction, which is a process to remove objects that are not considered to be parts of the image. Here, objects are regarded as noise when they have height and width between 1 and 7. These values are obtained based on a study indicating that characters have a minimum height and width of 8, therefore, objects with height or width below this minimum value can be considered as noise.

Manuscript segmentation initially finds the rows of script forming a manuscript, a process termed row segmentation. Then, it proceeds by isolating the script forming each row, called script segmentation. To conduct a rough segmentation on a row, the first tool used is the vertical projection profile, which projects the object pixels vertically to obtain information on which rows consist of groups of pixels forming the objects [11]. Because the script rows sometimes touch each other in manuscripts, the vertical projection results must be refined, for example, by moving average algorithm by Efstathiou, to obtain a curve that reflects the clarity of the distance between the phases relatively well. The phases in the curve are clues as to where a row starts or ends. Each phase shows one row of the script image.

Characteristically, Javanese manuscripts often do not show clear distances between rows; consequently, the trimming operation on the row image using the smoothed vertical projection will likely result in some scripts being trimmed as well. Each script that is entirely situated in the found row range will be clearly available; however, more often, parts of the scripts bleed over into the next row. To overcome this problem, the concept of pixel connectivity is applied to find the complete scripts. To address unreasonable row heights, the model used here first checks whether the row height is reasonable. Based on Javanese script characteristics, it is concluded that a reasonable row height equals the average height of all objects forming the manuscript plus two times of the standard deviation of the average height. The average height of the objects

forming the manuscript is obtained by evaluating all the objects in the manuscript using a connectivity operation between the pixels in any given object.

Each row image resulting from row segmentation becomes the input for the script segmentation process, which obtains the script images from the related row. Horizontal projection [11] is applied to obtain information concerning in which columns a group of pixels forming a script image are found. When the results of script segmentation can still be considered as a group of two or more scripts, a further process for script segmentation must be conducted. To find an initial clue about where the script should be segmented, the script width is evaluated. A script width is considered reasonable when it is no greater than the average width of all the objects forming the manuscript.

3.2. Script recognition

Considering inconsistencies in the result of handwriting, it is necessary to first improve the script image during the transliteration process to minimize writing differences. The transliteration model for Javanese script images is shown in Figure 3. The features of each script image to be transliterated are made uniform in terms of slant, thickness, and size. In this research, we use `maketform()` and `imtransform()` to repair the different slants. The Rosenfelt algorithm is implemented for reduction of the difference in thickness of scripts [12], and method `imresize()` is used to repair the difference in script size. From the result of studying several sizes of script image samples, we conclude that the minimum script image size of 88×88 pixels contains all objects in the script image.

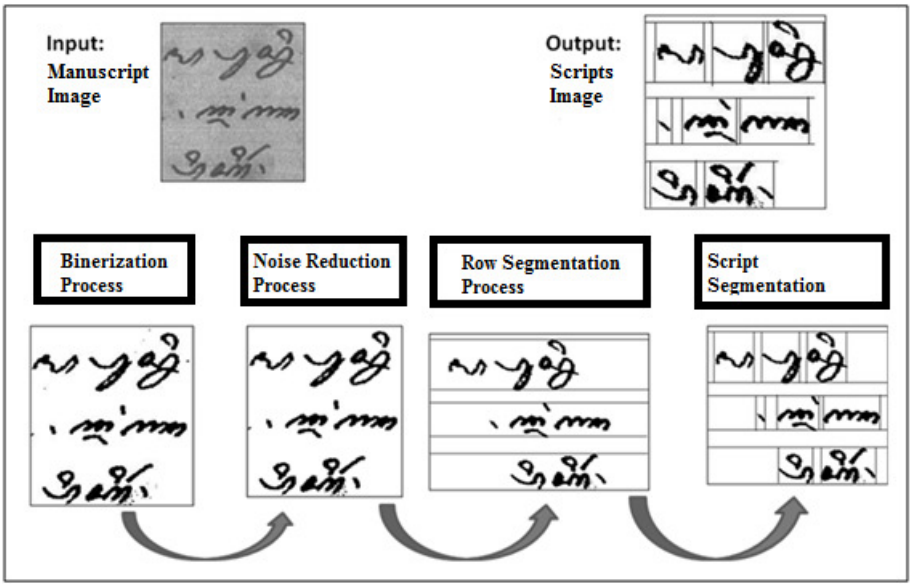

Figure 3. The model for recognition of Javanese script.

After obtaining uniformity in the external attributes of the Javanese script images that will be subjected to the transliteration process, the next stage is to find features that can be used as standards to determine the script image replacements. Surinto [13] proposed a method of Thai script image feature extraction, that can be replicated for Javanese script images. The proposed method involves counting the number of black pixels in certain areas of script.

In the process of classification, the choice of Latin name description or syllable is made by comparing the similarity between the input script's feature and feature of the image stored in a database. The method 1-NN is implemented for classification on this research.

3.3. Grouping syllables

The last process in transliteration is grouping syllables produced by the script recognition step into suite of words. The problem on this step is that there is no symbol for space in Javanese script, whereas a space is used to separate words. In this step we use dictionary as a tool for checking the word resulted from the last process. When a formed word occurs in the dictionary, then the syllable grouping is assumed to be correct. One problem that must be addressed is how many syllables should be grouped to form a single word. Widiarti and Winarko [14] reported that up to 7 Javanese language syllables can be grouped during word formation.

Assuming that all words that can possibly appear from a grouping of syllables resulting from the script image transliteration are available in the dictionary, the implementation of the grouping model for rows of syllables will certainly result in the appearance of rows of suitable words from the dictionary. However, previous research found that the transliteration model established for manuscript word recognition results in some errors, which are primarily due to incorrect script segmentation. These errors increase the probability for incorrect transliterations because they cause errors in subsequent feature extraction and script grouping. When such problems occur the syllable grouping will likely be unable to find a suitable matching word or will result in an incorrect match. Based on the probability that such problems will occur, another tool, in addition to the dictionary, is needed to check the formed words; one that predicts the appearance of a syllable based on previous syllables. Widiarti and Pulungan [15] proposed to use discrete-time Markov chains to predict the syllable that will appear next based on the highest probability of syllable occurrences after a previous specific syllable. They also proposed an idea of using a simple 2-dimensional matrix to manage the data for this syllable-probability table.

Based on the patterns of syllable groupings as described above, a model of syllable grouping is made and depicted in Figure 4. The process of grouping the syllables begins with Syllable_Correction, a stage that prepares the syllables before they are combined with other syllables to form words. After rows of syllables have been produced and are ready for further processing, the syllable-grouping stage begins. Each time a new word is produced by the syllable grouping process, the new word is searched in the dictionary. When any problem occurs, the probability table of the appearance of syllables is used as a tool to predict the appearances of other syllables when no word containing the syllable can be found. The selection of the syllable that follows is based on the syllables with the greatest appearance probability.

4. Evaluation of the method

The first evaluation is to analyse the method of manuscript segmentation. The primary data source used in this evaluation is a digitized copy of the manuscript with the catalogue number SB 141. Figure 5 shows a part of those manuscript, which has a *mbata sarimbag* style and stored in the Sonobudoyo Indonesia museum. The automatic segmentation of a portion of the SB 141 manuscript resulted in 291 scripts. Experts were consulted to determine which scripts were correctly segmented and which were not. Furthermore, 168 scripts were randomly selected to evaluate the segmentation result. The number of correctly segmented characters were counted, and the result showed that 91.07% of the scripts were correctly segmented. The same operation was repeated four times, and the obtained validity percentages were 89.88%, 92.86%, 89.29% and 88.69%, respectively. Based on these 5 tests, the average segmentation validity is 90.36%. Using this average transliteration success percentage of 90.36% and a confidence level of 95%, it can be concluded that the average validity of all manuscript segmentations using the proposed script segmentation model will be between 85.9% and 94.82%.

To obtain the range of the validity percentage of the transliteration model, an additional internal proportional test was conducted. There are 291 original Javanese script images segmented from the manuscript used to test the model. Using Slovin's formula, the minimum number of samples required for a statistical test is 168. To obtain an accurate result, 4 series

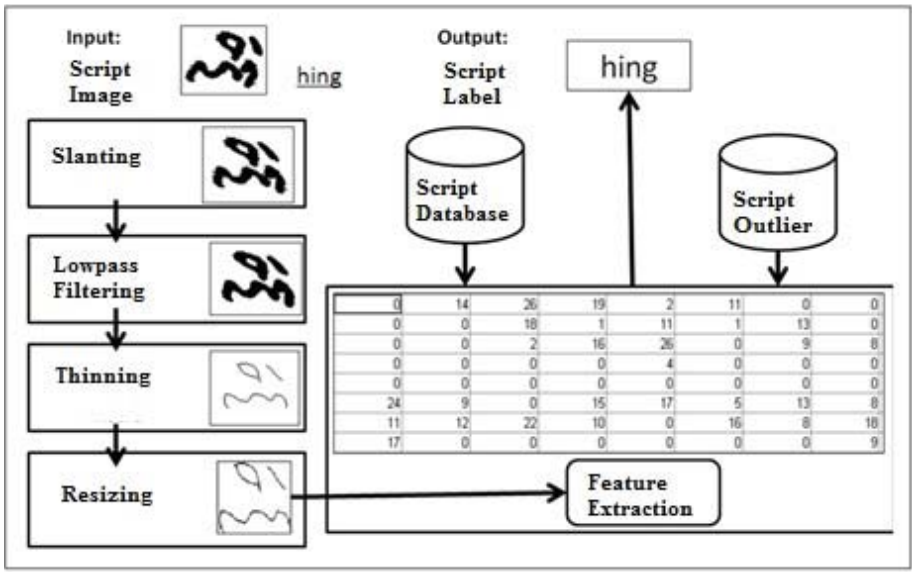

Figure 4. The model for syllables grouping from manuscript transliteration.

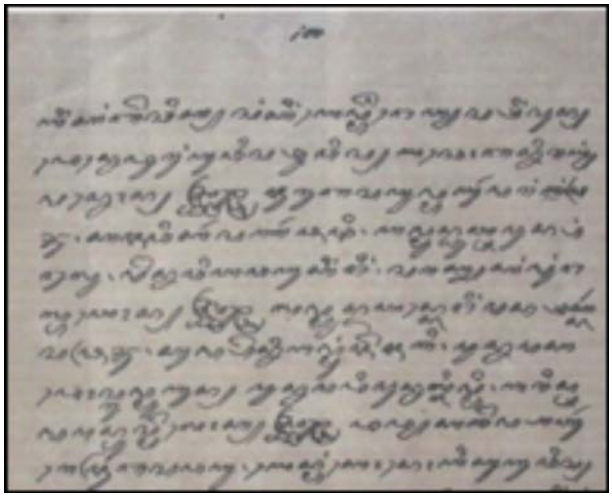

Figure 5. A part of SB 141 manuscript image.

of tests were performed similarly to the above, obtaining transliteration validity percentages of 81.2%, 79.2%, 78.6%, and 79.8%, respectively. Using these validity percentages, the new average validity percentage is 79.6%. With an average transliteration success percentage of 79.6%, a confidence level of 95%, and 168 data points, the confidence interval value of the average validity percentage of the transliteration falls between 73.51% and 85.69%. An analysis of one of the incorrect transliterations, indicates that the error is because there are script images that should be written separately, but instead, they were written in a connected manner.

The input data for testing the model of syllable grouping is a list of syllables transliterated automatically from the previous model test. The data for model testing may contain errors. Error in the test data may stem from errors that occur from the segmentation to the script transliteration phases. One manuscript used as test data has 100 known words marked by numbers to illustrate their order of appearances. The number 100 is taken as the population. Slovin's formula for a population of 100 words requires 80 words as the minimum sample size. By

using an analysis similar to the interval proportion test for the segmentation and transliteration models, experts first performed manual transliterations, and then 5 random selections were made from the 80 words. A check was made to determine whether any word in the order was produced correctly or not. Word production is assumed to be correct if it is similar to the manual transliteration conducted by experts and wrong otherwise.

In the first 80 words result of the transliteration, depicted in Figure 6, only 63 words were similar between the manual and system transliterations. Based on the correctly produced words from the total sample, the validity percentage is 80%. This procedure was repeated 4 times, achieving correct syllable groupings of 75%, 78.25%, 78.75%, and 78.75%, respectively. The average of these 5 validity percentages is 78.25%. Using the internal proportion test, it can be assumed that the average confidence interval of the validity percentage at a confidence level of 95% is between 69.20% and 87.29%.

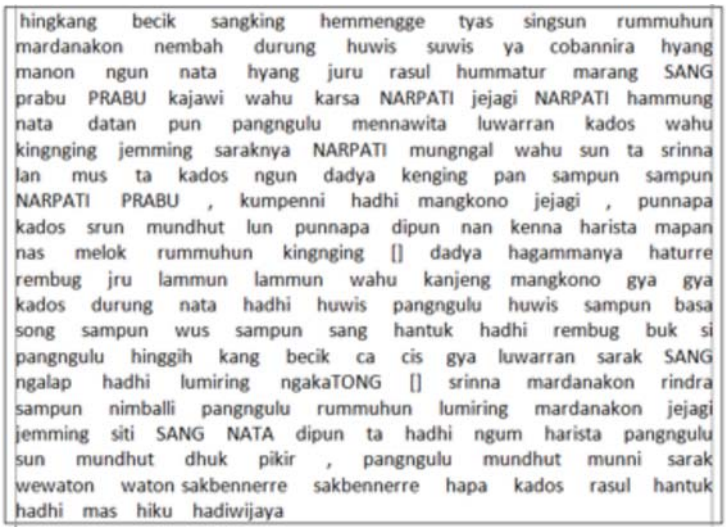
hingkang becik sangking hemmengge tyas singsun rummuhun
mardanakon nembah durung huwis suwis ya cobannira hyang
manon ngun nata hyang juru rasul hummatur marang SANG
prabu PRABU kajawi wahu karsa NARPATI jejagi NARPATI hammung
nata datan pun pangngulu mennawita luwarran kados wahu
kingnging jemming saraknya NARPATI mungngal wahu sun ta srinna
lan mus ta kados ngun dadya kenging pan sampun sampun
NARPATI PRABU , kumpenni hadhi mangkono jejagi , punnapa
kados srun mundhut lun punnapa dipun nan kenna harista mapan
nas melok rummuhun kingnging [] dadya hagammanya haturre
rembug jru lammun lammun wahu kanjeng mangkono gya gya
kados durung nata hadhi huwis pangngulu huwis sampun basa
song sampun wus sampun sang hantuk hadhi rembug buk si
pangngulu hinggih kang becik ca cis gya luwarran sarak SANG
ngalap hadhi lumiring ngakaTONG [] srinna mardanakon rindra
sampun nimballi pangngulu rummuhun lumiring mardanakon jejagi
jemming siti SANG NATA dipun ta hadhi ngum harista pangngulu
sun mundhut dhuk pikir , pangngulu mundhut munni sarak
wewaton waton sakbennerre sakbennerre hapa kados rasul hantuk
hadhi mas hiku hadiwijaya

Figure 6. The output of SB 141 manuscript transliteration.

5. Conclusion and future work

Javanese manuscript image transliteration involves replacing a row of images of Javanese script with a row of Latin text. The manuscript image transliteration model includes three interrelated and sequential submodels: manuscript image segmentation, Javanese script image to Latin script transliteration, and syllable grouping or word formation based on the transliterated Javanese script images.

Given the diversity in the height and width of Javanese script images written by the original manuscript authors, an adaptive approach becomes a non-negotiable rational approach to solve the problems in Javanese script image segmentation. The result of this research to solve the problems of Javanese script image to Latin script transliteration indicates that the structural approach can be applied to Javanese script feature extraction, with the requirement that the images of Javanese script that will be processed have been made uniform in certain physical characteristics: namely equal in size, thickness and slant. A dictionary of Javanese words is a reliable tool to test whether the result of syllable grouping from the Javanese script image transliterations forms valid words. To solve errors that occur while replacing the scripts, namely those that are due to the segmentation failures or errors in the script replacement process itself, a tool based on Markov chains is developed to predict the probability of the appearance of a syllable or a word based on previous syllables or words.

Acknowledgments
This research is funded by the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia, Penelitian Pasca Doktor scheme, grant no. 051/HB-LIT/IV/2017 and 010/HB-LIT/II/2018.

References

[1] Marsono 2010 *Centhini Tambangraras-Amongraga Jilid IV* (Yogyakarta: Gadjah Mada University Press)
[2] Sukmawati D L 2011 Inventarisasi naskah lama Madura *Manuskripta* **1** 17–30
[3] Behrend T E 1990 *Katalog Induk Naskah-Naskah Nusantara Jilid I Museum Sonobudaya Yogyakarta* (Jakarta: Djambatan)
[4] Behrend T E and Pudjiastuti D 1997 *Katalog Induk Naskah-Naskah Nusantara Jilid 3-A Fakultas Sastra Universitas Indonesia* (Jakarta: Yayasan Obor Indonesia)
[5] Budiarti E, Adji S E P, Setyawati K and Rahayu Y A 2007 *Karas Jejak-jejak Perjalanan Keilmiahan Zoetmulder* (Yogyakarta: Penerbit Universitas Sanata Dharma)
[6] Syahrul N 2011 Upaya dan penyelamatan naskah kuno Lampung *Manuskripta* **1** 1–15
[7] Baried S B, Soeratno S C, Sawoe, Sutrisno S and Syakir M 1985 *Pengantar Teori Filologi* (Jakarta: Pusat Pembinaan dan Pengembangan Bahasa Departemen Pendidikan dan Kebudayaan)
[8] Shridhar M and Kimura F 2000 *Handbook of character recognition and document image analysis* ed H Bunke and P S P Wang (Singapore: World Scientific Publishing Co. Pte. Ltd.) p 123–155
[9] Tangwongsan S and Sumetphong C 2008 Optical character recognition techniques for restoration of Thai historical documents *Proc. Int. Conf. on Computer and Electrical Engineering* (Thailand: IEEE Computer Society) 531–535
[10] Widiarti A R 2015 *Model Transliterasi Otomatis Citra Naskah Aksara Jawa* Dissertation (Yogyakarta: Universitas Gadjah Mada)
[11] Zramdini A and Ingold R 1993 Optical font recognition from projection profiles *Electronic Publishing* **6** 249–260
[12] Widiarti A R 2011 Comparing Hilditch, Rosenfeld, Zhang-Suen, and Nagendraprasad-Wang-Gupta thinning algorithms for Javanese character image *Proc. Int. Conf. on Pattern Recognition and Computer Vision* (Amsterdam: Academic Science Research) **57** 938-942.
[13] Surinta O and Schomaker L 2010 Overview of handwritten Thai character recognition *Lecture Notes Online* URL `http://www.ai.rug.nl/~mrolarik/APSMeeting/09-07-2010\%20Overview\%20of\%20Handwritten\%20Thai\%20Character\%20Recognition.pdf`
[14] Widiarti A R and Winarko E 2012 Algorithm for grouping syllables result from the Javanese literature document image recognition *Proc. of the 16th WSEAS Int. Conf. on Computers (part of CSCC '12)* (Greece: WSEAS) 266–271
[15] Widiarti A R and Pulungan R 2012 Aplikasi DTMC untuk post-processing pengenalan citra dokumen teks *Proc. Konferensi Nasional Sistem Informasi (KNSI 2012)* (Bali: STMIK-STIKOM) 75–78