Text block identification in restoration process of Javanese script damage

A R Himamunanto* and E Setyowati

Faculty of Science and Computer, Universitas Kristen Immanuel, Daerah Istimewa Yogyakarta, Indonesia

*The corresponding authors’ email address: rudatyo@ukrimuniversity.ac.id

Abstract. Generally, in a sheet of documents there are two objects of information, namely text and image. A text block area in the sheet of manuscript is a vital object because the restoration process would be done only in this object. Text block or text area identification becomes an important step before. This paper describes the steps leading to the restoration of Java script destruction. The process stages are: pre-processing, identification of text block, segmentation, damage identification, restoration. The test result based on the input manuscript “Hamong Tani” show that the system works with a success rate of 82.07%

1. Introduction

This research is one of a series preservation efforts of cultural heritage in the land of Java based on the ancient manuscripts. Manuscripts are literary works of the past and written in media such as leaves, animal skin, stone, wood, metal and so forth. Most of the manuscripts are in the form of writing, some of which are written with drawings as illustrations.

The information contained in manuscripts that are text and image generally have their respective roles. Therefore, it is necessary to do identification process that is able to separate between the object area of text and the object area of image to facilitate the next step such as text interpretation [1].

Digital image segmentation is an attempt to separate the components of visual objects that exist in a digital image [2]. Segmentation process in the manuscript input of Javanese script is expected to provide information about boundary in every object characters in the input of image document text.

2. Methods

Manuscript is an ancient literary written in various media such as: palm leaves, bamboo stems, paper, goat skin and others. Manual transliteration from Javanese script to be Latin script was done so the next generation will be understanding the history of the nation. There are large numbers of literary heritage only in Indonesia make the effort of transliteration process manually as an eternal task.

Involvement in the field of information technology and computer science is expected to support the Javanese literary experts in performing the transliteration of manuscripts. Manually transliteration process done by the Javanese philologist is considered to have been able to overcome all the conditions of damage on the document sheet Javanese script. Steps and action will be different with the automatically process on digital transliteration.

When a document image is received by a transliteration system, one of the basic things to do is to ensure that the input of the transliteration process is purely a Javanese script character. In fact, not all
of the input sheet of Java literary document contains only text block. This indicates that identifying the text area needs to be done before the next process [3]. It will separate between the text area and the image illustration.

Here is an example of a sheet from an ancient Javanese document entitled "HamongTani" by F S Winter exactly on page 56 that published by the publisher Batavia: Landsdrukkerij in 1877 as addressed in [4].

![Figure 1: Image document text page 56 from the book titled "HamongTani"](image)

2.1. Study of Javanese Script Writing

Javanese scripts consist of two major script groups, which are basic Javanese scripts and derivative Javanese scripts. Basic Javanese scripts are main Javanese scripts which have not been added with various punctuation marks or sandangan, therefore these main Javanese scripts are called legena or wuda scripts which means bare scripts.

![Figure 2: Nglegeno Javanese Script](image)

Generally, most Javanese scripts used do not only consist of legena scripts, but use various additions sandangan. There are many kinds of sandangan, i.e. sandanganswara for i consonant called wulu and is written above corresponding legena script, or sandanganswara of u consonant called suku which is written beneath legena script.

By 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 Figure 3.

![Figure 3: zone of Javanese script writing](image)
Area 1 is upper area or upper zone. In this zone there are sandanganswara i.e. wulu and pepet, also sandanganpanyigeg i.e. layar and cecak. Area 2 is main area or main zone. The area where basic scripts of Javanese scripts (legena script) are there. Area 3 is lower are or lower zone. Lower zone is a place for suku, lower part of taling, cakramandaswara script, cakrakeret, as well as pasangan placed below. Figure 4. shows an examples of Javanese scripts and places to put the scripts.

![Figure 4: Zone of Javanese script](image)

2.2. Type of Damage in The Javanese Script Document
The process of digitizing provides an opportunity to retain information as far as it can be done. Old manuscripts are generally over the age of 100 years have a variety of conditions due to time travel. In fact, even some manuscripts have been subjected to digital processes under various conditions. Some are caused by the vague ink, some of the character lines that have been blurred until some media that has been torn and lost. One example of visual impairment is the reduced thickness of the script line and it results in the loss of the edges of the script line. Some damage cases even result in a broken character line. Some examples of damage are shown in Figure 5.

![Figure 5: some examples of Javanese script damage](image)

For some people who are quite used to using Javanese language or at least familiar with Javanese script assume that the visual damage of Javanese script characters can still be read aksaranya. Problems may arise when the image of the Javanese script document will be subject to advanced image processing, for example for the purposes of Javanese script transliteration [5]. The restoration of Javanese script image becomes an important step in order to get the image of the damaged Java script getting visual improvements. Visual impairment of a script image raises multiple interpretations for the reader or for further processing.

2.3. Proposed Steps
The stages of digital restoration process in this research involves 5 main processes namely: (a) image acquisition, (b) preprocessing, (c) identification the text area (ITA) of Javanese script, (d) character segmentation, (e) damage identification and (f) restoration of script. Pre-processing is an initial process where in the input image to done the process such as: grayscalle operation, binary operation, noise reduction operations. ITA is a process used to identify and mark only for the area text of Javanese script. Character segmentation is a process used to separate objects of the script with each other. The output of this process is an object component of Javanese script. Damage identification is is a process used to check whether an object is damaged or not. Checking is done by running skeletoning operation and get an extraction feature of the object script. Statitical approach is the algorithm used to got the feature or template [5]. Template matching is process to ensure that the corresponding characters are not damaged. Template matching base on statistical approch is a method that used. If based on the characteristics obtained is known that a character is declared damaged then the restoration process activated.
3. Result and Discussion

3.1. Shape of The Javanese Script Damage

The ancient manuscripts chosen as inputs in this experiment are the manuscript entitled "Hamong Tani" written by Karel Frederik Holle in 1874-1879 and translated to the Javanese script by F.S. Winter in 1876-1878 and published by the publisher Batavia: Landsdrukkerij in 1877.

Here is the character damage Javanese script that has been grouped based on the level of damage to the test book titled "Hamong Tani". The damage that is classified as level 1 is the damage caused by the loss of black dots on the edges of the script line in small numbers. It is shown in Table 1.

| Shape of damage | Name |
|-----------------|------|
| Ho              |      |
| Co              |      |
| Ko              |      |

Table 1. Example of damage at level 1

Damage classified level 2 is the damage caused by the loss of black dots on the edge side of the line of characters in a larger number and potentially break the line of character. As shown in Table 2.

| Shape of damage | Name |
|-----------------|------|
| Ho              |      |
| Co              |      |
| Ko              |      |

Table 2. Example of damage at level 2

Damage classified as level 3 is the damage due to loss of black dots and break the line of script. As shown in Table 3.

| Shape of damage | Name |
|-----------------|------|
| Ho              |      |
| Co              |      |
| Ko              |      |

Table 3. Example of damage at level 3

3.2. Identify Text Block

Identify Text Block (ITB) is a process undertaken to identify the existence of Javanese script text blocks and grouped separately from the illustrative drawing area.

Figure 6: (a)(b)(c) are results of step by step in ITB
3.3. Character Segmentation

The input of these processes is the output from identifying the text block (ITB) that has been done before. Only the inputs are text objects that will be processed in Java character object segmentation. This process is done to obtain independently character by character from the script. The method used to segmenting each character object is projection profile. Figure 7 below is the result of the segmentation process on input of Javanese script as discussed in [2].

![Figure 7: Results of segmentation with vertical projection](image)

3.4. Damage Identification

Damage identification is the process to checking each Javanese character and ensure that a character is not visually damaged. The input of this process came from the character segmentation process. The damage identification process involves pattern recognition operations. The character pattern result then compared with the true character pattern stored in the database. The equal result means the character does not experience damage but if the comparison show different values then the restoration process is done. Restoration is done by replacing the correct characters.

![Figure 8: (a) origin image, (b) result of thinning and (c) feature extraction result](image)

3.5. Result Testing System

The results of the testing process stage are the nearest script image which is expected to give the best chance to be the image of the candidate or the image of the restoration result. Some examples of test results are as presented in the following table.
Table 4. Part of the test result

Based on the test results with as many as 452 Java script inputs that suffered a variety of damage, it is known the number of characters that otherwise found the correct restoration results are as follows:

\[
\text{percentage of truth} = \frac{\sum \text{truechar}}{\sum \text{allchar}} \times 100\% = \frac{371}{452} \times 100\% = 82.07\%.
\]

4. Conclusion

The proposed of text identification have been tested on all of pages printed Hamong Tani documents and works not bad. Base on testing shows that the level success is 83.07%. The unidentified text area known to be manually it is probably caused step in identification of text area and segmentation needs to evaluate.

5. References

[1] Santosh J and Livingston L M 2013 Text detection from documented image using image segmentation International Journal of Technology Enhancements and Emerging Engineering Research 1 4.

[2] Himamunanto A R and Widiarti A R 2013 Javanese character image segmentation of document image of Hamong Tani Digital Heritage International Congress (DigitalHeritage) 1 641-644.

[3] Chen D, Bourlard H and Thiran J P 2001 Text identification in complex background using SVM Computer Vision and Pattern Recognition 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on 2 II-II.

[4] Winter W S 1876 Hamong Tani (Batavia: Landsdrukkerij).

[5] Himamunanto A R and Widiarti A R 2015 The Automatic Transliteration Model of Images of the Book Hamong Tani Using Statistical Approach World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering 2 8.

Acknowledgments

This research is fully supported by 2017 research grant of Indonesian Higher Education Ministry (RISTEKDIKTI)