Digitization of Batak Manuscripts Using Methods Learning Vector Quantization (LVQ)

M A Muchtar¹, I Jaya¹, Nandar Cholid Siregar¹, E B Nababan¹, Syahril Effendi², O S Sitompul¹, M Zarlis², U Andayani¹, T H Nasution³, I Siregar⁴, M F Syahputra¹

¹Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara
²Department of Computer Science, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara
³ Department of Electrical Engineering, Faculty of Technic, Universitas Sumatera Utara
⁴ Department of Industrial Engineering, Faculty of Technic, Universitas Sumatera Utara

Email : angi.muchtar@usu.ac.id.

Abstract. Digitalization of the Batak manuscript itself is still done manually. Dr. Uli Kozok has made a Batak script font that can run on Windows, Linux, and Apple OSX but there is no system that can digitize Batak manuscripts properly. Therefore, we need a system that can digitize Batak manuscripts. This system was built by applying the LVQ Method. Batak manuscript digitization in .jpg format is used as an image input file for the recognition process. Then image pre-processing including grayscale, contrast, thresholding, and segmentation is done in order to facilitate the recognition of Batak documents and recognition as a result of document recognition. The application of LVQ method can digitize the Batak manuscript well and in accordance with what is desired. This is supported by an accuracy rate of 97.9%. Based on system testing, the absence of training data in the input document, the letters are unclear, the position of the letters is not appropriate and the background in the document is still noise which greatly affects the success rate of the system in recognizing Batak script documents.

1. Introduction

Batak script is one of the diversity of arts and culture in Indonesia. The culture and arts are dance, language, music and many more and the Batak tribe consists of several tribes namely Karo, Mandailing, Toba, Simalungun, and Pakpak. If seen from its history, the Batak script is an Indian writing family. The oldest Indian script is the Brahmi writings which revealed two groups namely North India and South India. Batak script can also be classified as an abugida (a type of phonetic writing whose language sounds can be accurately represented). There are two types of Batak alphabet sets, namely: “ina ni surat” (main letters) and “anak ni surat” (derivative letters). Indonesia is also an archipelago that has a diverse cultural heritage, one of which is manifested in the scripts or original writings of the regions included in the archipelago script. Batak script is a family of Batak scripts that needs full attention due to limited data and information. The digitalization of the Batak manuscript has so far been carried out manually. Therefore, the application of a technology is needed to be able to digitize Batak manuscripts.

Previous research has been conducted, about character recognition between Myanmar and English characters. The system will change the Portable Document Format (.pdf) file to Machine
Editable Word Document (.doc). The introduction of Myanmar and English characters will use MICR (Myanmar Intelligent Character Recognition), which is one part of OCR. 92% success rate for the introduction of Myanmar and English scripts.

There is also research carried out, [5] on Kannada's Handwriting Character Recognition Using Hybrid Features. In this study, the hybrid feature extraction technique was used to recognize Kannada's handwriting. Kannada's handwritten characters are written in different font sizes and styles. The features taken from each image are 3600 samples that are used as training datasets to get good feature scores. The pre-processing stage has several steps to implement the hybrid feature method, namely correction, binarization, and noise removal. And the resulting output is in the form of documents containing parts of text then printed in the form of handwriting. Through this approach, researchers obtain an accuracy rate of 87.33%.

Subsequent research was conducted, [7] on the Introduction to Digit Handwriting Using LVQ. In this paper 10 (ten) structural features are used for handwriting recognition. Digit recognition system depends on the type of features used and the data used on this system as many as 50000 training data images and 5000 data testing. The success rate on this system reaches 95%.

The following research was conducted, [6] a comparison of the handwriting recognition of Arabic characters with Arabic characters that have been printed using the discrete method Hidden Markov Models (HMM). In this study, after the character pre-processing step, the segmentation process is continued using a recursive algorithm and feature extraction based on the structural properties of the text. The results of this study indicate an introduction rate of 93.40% for Arabic handwriting.

In addition, research has also been conducted, [3] on the Introduction of Arabic Handwriting Optical Characters using Artificial Neural Network. In this system Artificial Neural Network is used as a classification. Artificial Neural Networks are trained based on the Hopfield Algorithm. And the output of this system is able to recognize eight Arabic handwritten letters well. Through this approach, the accuracy rate on this system is 77.25% for Arabic handwriting.

Based on the description of the research that other researchers have done, the basic problem of digitizing the Batak manuscript itself is still done manually. There is no system that can digitize the Batak manuscript itself automatically. Therefore, we need a system that can digitize the Batak manuscript properly through the input image of the scanned .jpg image file and then convert it to font.docx.

2. Data and Methods Used

2.1. Data Used

The method proposed by the author for digitizing Batak manuscripts in this study uses LVQ. The image to be processed is in the form of an image with .jpg input which will then enter the image pre-processing process. In this process the image will undergo grayscale, contrast, and thresholding processes which will produce gray, black and white images to facilitate the next process. Then the segmentation process will be carried out, namely cropping, where each document will be cut in lines per line. The next process is the application of the LVQ method to digitize the Batak manuscripts, whether it is the Mandailing, Toba, Karo, Simalungun and Pakpak Batak scripts. The general architecture in digitizing Batak manuscripts in this study can be seen in figure 1.

2.2. Method Used

The general architecture in this study consists of several stages, namely:

2.2.1. Input

Training data is preliminary data that will be processed, while testing data is data that will be tested on the system and compared with previous data. At this stage, document files in .jpg format are included as testing data to be tested and processed by the system. The Batak script documents in this document file are script documents both from books, and bark with the font Dr. Uli Kozok.
2.2.2. Pre-processing
Image pre-processing is the initial stage in this research. This stage aims to process the image that is input and converted into an image that will be further processed. The stages in image pre-processing are as follows:

**Figure 1.** General architecture.

**Figure 2.** Greyscale process.

**Figure 3.** Contrast process.
2.2.2.1 Grayscale
Menurut Basuki (2005), the initial process that is mostly done in image processing is changing the color image to grayish (grayscale). The grayscale process in this study can be seen in figure 2.

2.2.2.2 Contrast
At this stage a sharper contrast is carried out in order to facilitate the system to proceed to the next stage. The contrast process in this study can be seen in figure 3.

2.2.2.3 Thresholding
A thresholding image is an image where each pixel is only expressed by a value of two possibilities (i.e. values of 0 and 1). A value of 0 represents black and a value of 1 indicates white. Then the function of thresholding images at this stage is to simplify the process of digitizing Batak manuscripts, because documents will be more easily detected on images that contain less colour. The thresholding process in this study can be seen in figure 4.

2.2.3 Image Segmentation
The images used for input are Karo, Toba, Pakpak, Simalungun and Mandailing documents. The size for all images is 24 x 42 pixels, because it is to facilitate the user in processing the image of the Batak script document itself. The input that will be processed at this stage is the image that is scanned or photographed, and the system automatically captures the image to be processed to the next stage. The following is an example of the image when doing the cropping process can be seen in figure 5.

2.2.4 Learning Vector Quantization
LVQ is a method for learning in supervised competitive layers. A competitive layer will automatically learn to classify input vectors. The classes obtained as a result of this competitive layer only depend on the distance between the input vectors. If two input vectors approach the same, the competitive layer will put both input vectors into the same class[8].

![Figure 4. Threshold image process.](image)

![Figure 5. Cropping results image.](image)
Explanation:

a) $x_1$ until $x_{19200} = \text{input value}$

b) $|| x - w_1 ||$ until $|| x - w_n || = \text{weight distance}$

c) $H_1$ until $H_n = \text{output layer}$

d) $D_1$ until $D_n = \text{output value}$

e) $n = \text{number of character data (a number of class)}$

where:

- $x_1$ until $x_{19200}$ is a matrix element in each character pattern that will be used as an input value.
- $|| x - w_1 ||$ until $|| x - w_n ||$ is the calculation of the smallest distance between $w_1$ and $w_n$ is the initialization data value.
- $H_1$ until $H_n$ are output layers.
- $D_1$ until $D_n$ is the final weight that will be used later in the testing process with new character data entered.

2.2.4.1 Learning Vector Quantization Learning Algorithm

The LVQ method will first make an introduction to the input patterns that will be presented in vector form so that the class can be searched. Because each output neuron represents a certain class or category, the input pattern can be recognized by the class based on the obtained neuron output. The LVQ method recognizes the input pattern based on the distance approach between two vectors, namely the vector of input units / neurons and the weight vector. The introduction of the LVQ method has two processes:

- Training process
- Testing process

In the initial training process, the input vector will experience a learning process that will be carried out through several epochs until the maximum epoch limit is reached. LVQ conducts learning in supervised competitive layers. A competitive layer will automatically learn to classify input vectors. The classes obtained as a result of this competitive layer only depend on the distance of the input vectors with the weight vectors of each class and the input vectors will enter the class that has the closest distance. The parameters used in this LVQ method are as follows:

1. Alfa (Learning rate) Alfa is defined as the level of learning. If the alpha is too large, then the algorithm will become unstable otherwise if the alpha is too small, the process will take too long. The alpha value is $0 < \alpha < 1$.
2. DecAlfa (Decreased Learning rate) That is a decrease in the level of learning.
3. MinAlfa (Minimum Learning rate) That is the minimum value of the level of learning that is still allowed.
4. Max Epoch (Maximum epoch); i.e. the maximum number of epochs or iterations that can be done during training. Iteration will be terminated if the value exceeds the maximum epoch.
2.3 Flowchart
Flowcharts are graphical descriptions of the steps and sequence of procedures of a program. Batak manuscript digitization system flowchart can be seen in figure 7.

![Flowchart](image)

**Figure 7. Batak manuscript digitalization flowchart.**

2.4 Output
In this process, the Batak document file that has been done step by step will produce a document font with the .docx extension which has been structured according to the document input data, and then will be stored on the computer.

3. Result and Discussion
Batak manuscript digitization system is a system that can make it easier for users to process images by going through several image pre-processing for better results. This Batak Manuscript digitization application starts with the home view, where in this page has menus that have their respective functions. Before making an introduction, the user first conducts training data with the aim of facilitating the user in conducting an introduction system as shown in figure 8. When the user will digitize the Batak manuscript, then the next step to do is to select the Recognition menu on the home page. The following operational procedures that users must carry out in digitizing the Batak manuscript system as shown in figure 9.
If the user has done training data as shown in figure 10, then the next step is Recognition. The user returns to the home page, then selects the Recognition menu, the system switches to Recognition and the user can choose what script they want to process as shown in figure 11. The recognition display can be seen in figure 12. After the user chooses Recognition precisely in the Karo character, then the recognition display, where the user will enter the Batak script document file that will be recognized by the system. Then the user can select the "Browse File" button which will direct the user to the storage.
of the document file and select the Batak script document to be recognized. At this stage, the author makes an example document that will be processed is document_karo1.jpg and click Open. The following dok_karo1 that will be processed to the next stage can be seen in figure 12. After the system displays the document file that you want to process, then the next step is pre-processing the image starting from grayscale, contrast, thresholding and recognition. Figure 13 can be seen the occurrence of image pre-processing. For grayscale processes marked in black, contrast is marked in red, and thresholding is marked in orange. After the user presses the recognition button, the system will automatically digitize the Batak manuscript by cropping rows by line until the operation completed message appears, can be seen in figure 14.

And this is the result of recognition that has been done by selecting the file dok_karo1. The following is the result of the dok_karo1 file recognition with the name "dok_karo.docx" can be seen in figure 15. After the recognition results are complete, then the dok_karo.docx file will automatically be saved in the file storage directory. System testing is done to find out whether the system that has been built is functioning properly and runs as desired. This test will explain how much the level of accuracy in the implementation of LVQ is applied to the Batak manuscript digitization system. This testing process uses data that has been obtained in previous studies. The test results on this system will use the formula:

\[
\text{Accuracy} = \frac{\text{Total True Testing Data}}{\text{Total All of Data}} \times 100\%
\]

Testing the Batak manuscript digitization system and calculating the accuracy of the fit of the test results of each Batak script document can be seen in Table 1.

| No. | input document | successfully recognized | Amount of manuscripts | Index error | Accuracy |
|-----|----------------|------------------------|-----------------------|-------------|----------|
| 1   | dok_karo1      | 225                    | 228                   | 3           | 98.6%    |
| 2   | dok_karo2      | 258                    | 211                   | 3           | 98.5%    |
| 3   | dok_karo3      | 216                    | 216                   | 0           | 100%     |
| 4   | dok_karo4      | 217                    | 219                   | 2           | 96%      |
| 5   | dok_karo5      | 264                    | 211                   | 7           | 94.9%    |
| 6   | dok_karo6      | 200                    | 212                   | 4           | 98.1%    |
| 7   | dok_karo7      | 216                    | 210                   | 3           | 98.6%    |
| 8   | dok_karo8      | 220                    | 220                   | 0           | 100%     |
| 9   | dok_karo9      | 215                    | 217                   | 4           | 98.1%    |
| 10  | dok_karo10     | 216                    | 214                   | 0           | 100%     |
| 11  | dok_karo11     | 219                    | 210                   | 0           | 100%     |
| 12  | dok_karo12     | 215                    | 215                   | 0           | 100%     |
| 13  | dok_karo13     | 216                    | 216                   | 0           | 100%     |
| 14  | dok_karo14     | 208                    | 213                   | 5           | 97.6%    |
| 15  | dok_simahun1   | 217                    | 220                   | 0           | 98.6%    |
| 16  | dok_simahun2   | 219                    | 221                   | 2           | 99%      |
| 17  | dok_simahun3   | 225                    | 228                   | 3           | 98.6%    |
| 18  | dok_simahun4   | 225                    | 227                   | 4           | 98.2%    |
| 19  | dok_simahun5   | 250                    | 216                   | 16          | 92.5%    |
| 20  | dok_manailing1 | 225                    | 231                   | 5           | 97.8%    |
| 21  | dok_manailing2 | 229                    | 234                   | 5           | 97.5%    |
| 22  | dok_manailing3 | 223                    | 220                   | 5           | 97.3%    |
| 23  | dok_manailing4 | 226                    | 232                   | 6           | 97.4%    |
| 24  | dok_manailing5 | 229                    | 233                   | 4           | 98.2%    |
| 25  | dok_manailing6 | 224                    | 227                   | 3           | 98.6%    |
| 26  | dok_karo15     | 119                    | 124                   | 14          | 98.7%    |
| 27  | dok_total2     | 45                     | 60                    | 15          | 75%      |

Jumlah      8864  8761  117  97.9%
As seen in Table 1, that dok_karo1 has an error index (unknown) of 3 letters. Examples of unknown letters can be seen in figure 16.

![Figure 16. Example Batak script input dok_karo1.](image)

In figure 16, there are 3 letters that are not recognized (error) that are marked in red, because the letters in the Batak script documents are not listed in the Karo training data and the letter positions are not appropriate. So, there is no training data in the input document, the type of letters is unclear, the position of the letters is not appropriate and the background on the document greatly influences the success of the system in recognizing the Batak script documents.

4. Conclusions

Based on the results of research conducted, it can be concluded that the digitization system of the Batak manuscript uses the LVQ method as follows:

1. The application of the LVQ method can recognize Batak script documents properly and in accordance with what is desired. This is supported by an accuracy rate of 98%.
2. Based on system testing, the background to the document, the position of the letters, the type of letters is unclear and the absence of training data in the input document greatly influences the success rate of the system in recognizing Batak script documents.
3. The success rate of the system in digitizing book-shaped Batak manuscripts is 98%. While the document in the form of wood bark by 75%.
4. There are still documents recognition errors. For example, if there is no training data in the input document, the typeface is not clear, the letter position is not appropriate and the background is still noise, then one of the letters in the input document will be an error (not recognized).

Referensi

[1] Yadana and Cherry M 2010 Mixed-Myanmar and English Character Recognition with Formatting. IJCSI International Journal of Computer Science 7(6).
[2] https://ulikozok.com/aksara batak/sejarah-aksara batak/
[3] Hussien S R 2015. Optical Character Recognition of Arabic Handwritten Characters using Neural Network. International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering.
[4] Kozok U 2017. Surat Batak . retrieved from : http://ulikozok.com
[5] Pasha S., Padma MC 2013. Recognition of Handwritten Kannada Character using Hybrid Features. Department of Information Science and Engineering, PES College of Engineering, Mandya - 571401, Karnataka, India
[6] Firoj P 2012. An Empirical Evaluation of Offline Arabic Handwriting And Printed Characters Recognition System. IJCSI International Journal of Computer Science 9(6),
[7] Ravi, U B., Venkateswarlu Y, Kumar C A. 2014. Handwritten Digit Recogniton Using K-Nearest Neighbour Classifier. Aacharha Nagarjuna University, Assoc, India.
[8] Rahmat R F, Pulungan A F, Faza S, Budiarto R 2017 Image Classification of Ribbed Smoked Sheet using Learning Vector Quantization Journal of Physics: Conference Series 801(1), 012050