Feature Extraction Comparison in Handwriting Recognition of Batak Toba Alphabet

Novie Theresia Br Pasaribu[1], M. Jimmy Hasugian[2]

***Abstract*—Offline handwriting recognition is one of the most prominent research topics due to its tremendous application and high variability as well. This paper covers the offline Batak Toba handwritten text recognition, from the noise removal, the process of feature extraction until the recognition by using several classifiers. Experiments show that elliptic fourier descriptor (EFD) is the most discriminative feature and Mahalanobis distance (MD) outperforms the two others classifier.**

***Keywords*— Batak Toba Alphabet, handwriting recognition, feture extraction, classification.**

I. Introduction

Batak Toba is one of the tribes in Indonesia that has its own language and script. However, not many people know about it, including the Batak Toba people. Batak Toba script is one of the Indonesian cultures that should be preserved. One of the proposed attempts is the recognition of handwriting and ancient manuscripts of the Batak Toba script.

The Batak Toba script includes Brahmi (Indian) writings, especially the Southern Indian writing group. The Batak Toba script can also be classified as an abugida, a phonetic text that any sound can be accurately denoted [1].

The Batak Toba script is dominated by a curved stroke, which has similarities with Arabic script. This is the motivation in choosing the various feature extraction methods used [2], [3].

On the other hand, handwriting recognition is one of the most important research topics to study because it has high variability, such as the tendency of different handwriting of each person and the background of writing that is sometimes not free from noise.

This paper focuses on the recognition of the handwriting of the Batak Toba script with various background noise on writing and with various feature extraction. Giving various kinds of noise on the background of writing aims to imitate the existing writing on ancient manuscripts of the Batak Toba. Handwriting recognition of the Batak Toba script is done with various feature extractions to see the exact feature extractions represent the Batak Toba script.

II. Batak Toba Alphabet

In the Batak Toba language, the Batak Toba script is often called *si sia-sia* or *surat sepuluh sia,* because the number of letters (*ina ni surat*) are nineteen [1]. In the Batak Toba language, *sampulu sia* means nineteen. The nineteen letters are shown in Fig. 1.

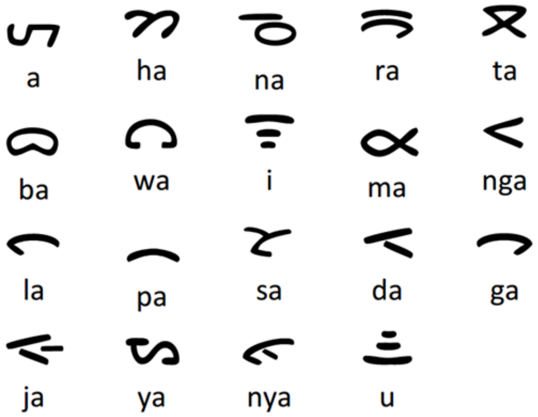


Fig. 1 The *ina ni surat* Batak Toba alphabets.

The above Batak Toba script is a modern version, which is still commonly found in present writings and incorporated into the learning curriculum as a local content in several schools in Toba Samosir District, North Sumatra Indonesia. Batak Toba script writing method is similar to Latin letters from left to right, which are consonants ending with [a]. The sound [a] attached to *ina ni surat* may be converted into another vowel by adding *anak ni surat* as shown in Fig. 2.

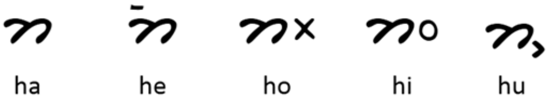


Fig. 2 Vocal sound changes because of *anak ni surat*.

The ancient manuscripts with the script of the Batak Toba were written with materials made of wood, bamboo, animal skin, or bone. One example of an ancient manuscript written on wood is shown in Fig. 3.

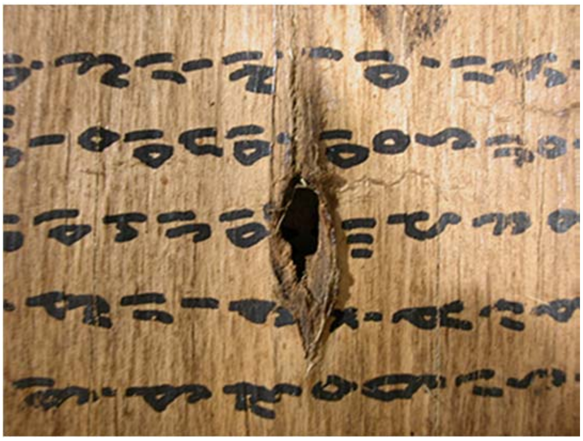


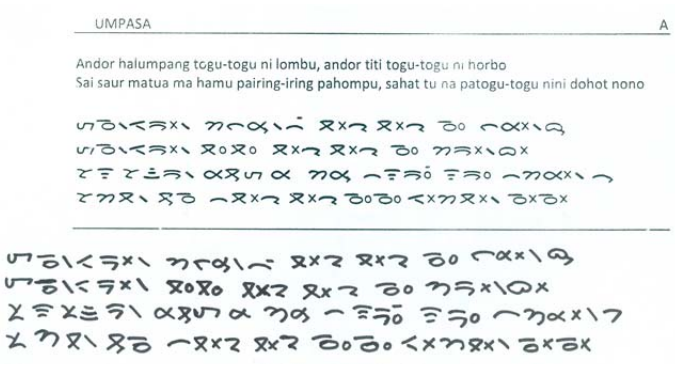
Fig. 3 Ancient manuscript of Batak Toba Alphabet written on wood (source: TB Silalahi Center, Balige, Sumatera Utara).

[1,2] *Lecturer, Program Studi Teknik Elektro, Universitas Kristen Maranatha, Jln Surya Sumantri no 65, Bandung, 40164, INDONESIA (e-mail: novie.theresia@eng.maranatha.edu, e-mail: jhasugian@maranatha.edu)*

III. Methodology

The handwriting of the Batak Toba script described in this paper was obtained from hundreds of students' handwriting at several schools in Balige, North Sumatra Indonesia. Students were asked to write one of the common six proverbs among the Batak Toba tribe on a piece of paper.

Then the results of handwriting was scanned with a resolution of 300 dpi. A handwriting example for a type A proverb is shown in Fig. 4.



UMPASA A

Andor halumpang togu-togu ni lombu, andor titi togu-togu ni horbo
Sai saur matua ma hamu pairing-iring pahompu, sahat tu na patogu-togu nini dohot nono

Fig. 4 Example of student's handwritten about Batak Toba poetry.

This handwritten collection has been documented and accessible for free [4]. The block diagram of the Batak Toba script handwriting recognition process is shown in Fig. 5.

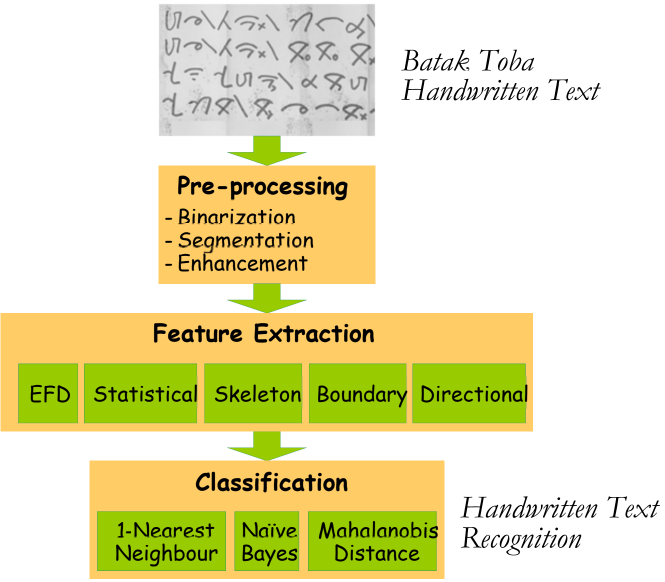


Fig. 5 Block diagram of Batak Toba handwritten recognition.

A. Pre-processing Stage

Before the pre-processing stage as shown in Fig. 5, the handwriting is scanned using two processes: (1) the binary process by using the threshold method as the target data in the cleaning process [5], (2) the scanned image is combined with several background types. The performed process is shown in Fig. 6.

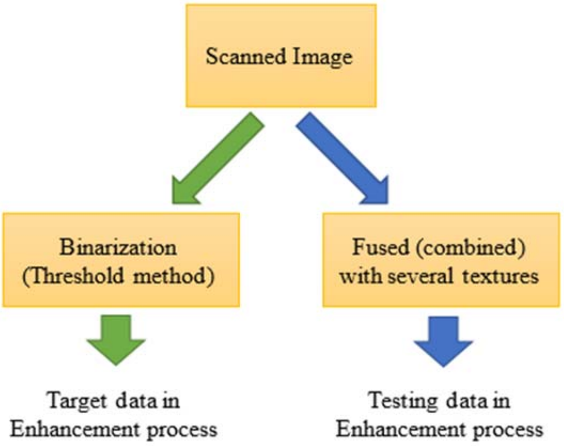

Scanned Image
Binarization (Threshold method)
Fused (combined) with several textures
Target data in Enhancement process
Testing data in Enhancement process

Fig. 6 Process of preparing data for the system.

Both processes were undertaken to provide data simulation for the handwriting recognition system.

Fig. 7 illustrates the process of combining the scanned text of handwritten $I_s(i,j)$ with a paper background $I_s(i,j)$. This process was done using the wavelet-based image fusion method to produce handwritten images that have noise [6]. This method is used because it produces a composite image (fusion) that gives the effect of visual perception resembling writing in ancient manuscripts. In the experiment, four types of paper background were used in the merging process as shown in Fig. 8.

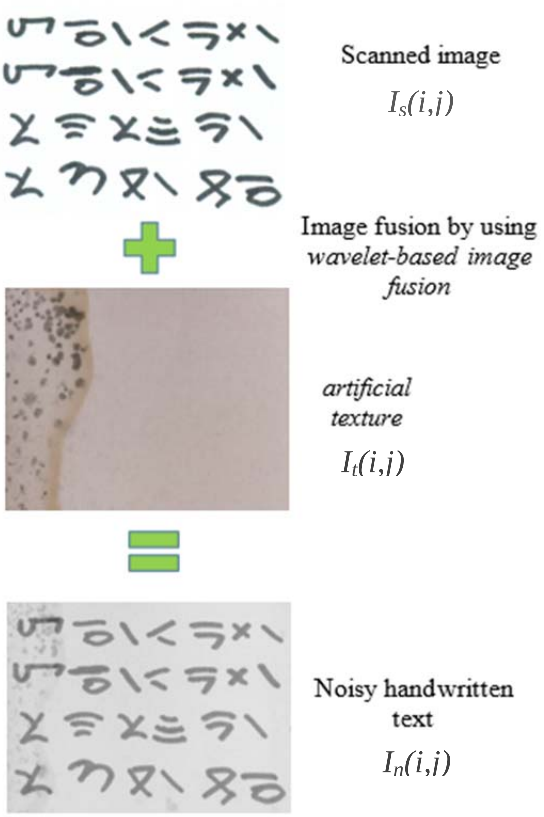


Fig. 7 The fusion of text and artificial texture.

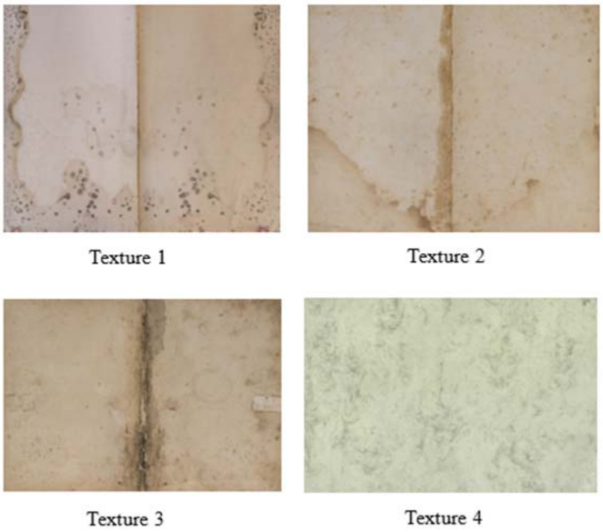


Fig. 8 Four artificial background textures used in the system.

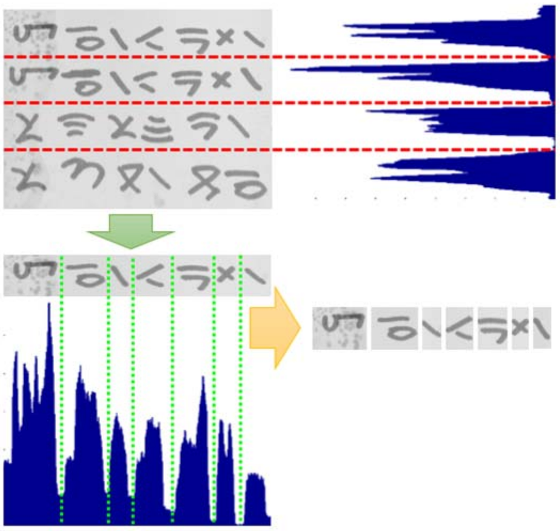


Fig. 9 Line and character segmentation process.

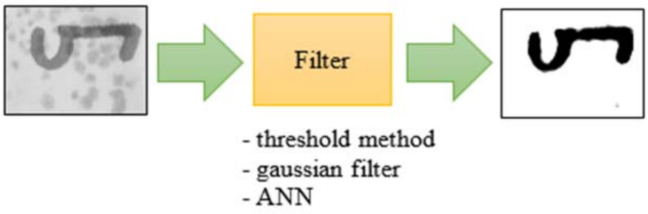


Fig. 10 Enhancement process of Batak Toba alphabet.

After creating handwritten images with the background of various types of textures, then the segmentation stage is performed.

The segmentation process was done using the horizontal projection method for the line segmentation, then the vertical projection method for segmentation per character [7]. The process diagram is shown in Fig. 9.

The process of enhancement, or denoising to remove noise due to background text was done with a schema as presented in Fig. 10.

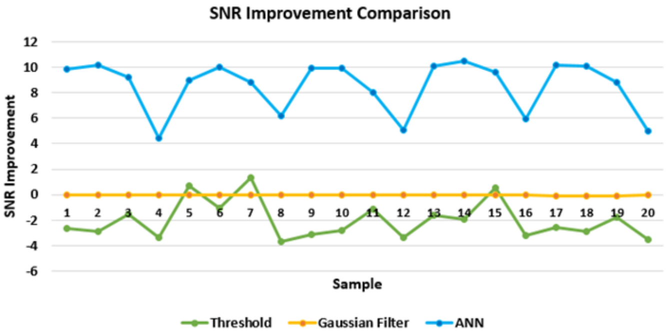


Fig. 11 SNR Improvement comparison of the three methods.

There are three methods compared in this process, namely threshold method [5], Gaussian filter, and Artificial Neural Network (ANN). The filter selection criteria were based on the SNR Improvement value as in (1).

$$\mathrm{SNR} = 10 \times \log_{10} \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}[I(i,j) - I_n(i,j)]^2}{\sum_{i=1}^{M}\sum_{j=1}^{N}[I(i,j) - I_d(i,j)]^2} \,. \quad (1)$$

$I(i,j)$ is a handwritten image that has undergone a binary process, such as Fig. 6, it is an image that has been merged with an artificial background (texture), such as Fig. 7, it is the output image of the filter.

From the results obtained earlier, the enhancement process with ANN shows SNR Improvement values that outperform two other filters as shown in Fig. 11 [8].

B. Feature Extraction Stage

At the feature extraction stage, there are 51 feature vectors of the five categories used, namely: (1) Statistical; (2) Skeleton; (3) Boundary; (4) Directional; and (5) Elliptic Fourier Descriptor (EFD) [2]. Details of the feature extraction are described as follows:

1) Statistical Features: There are 14 features in this category. The binary image area is calculated using (2).

$$A = \sum_{x}\sum_{y} B(x,y) \quad (2)$$

The image mass center $(\bar{x},\bar{y})$ is also used to calculate normalized central moments $\eta_{2,0}$ and $\eta_{0,2}$ through (3).

$$\eta_{u,v} = \frac{1}{A^k}\sum_{x}\sum_{y}(x-\bar{x})^u(y-\bar{y})^v B(x,y) \quad (3)$$

with $k = 1 + \frac{u+v}{2}$.

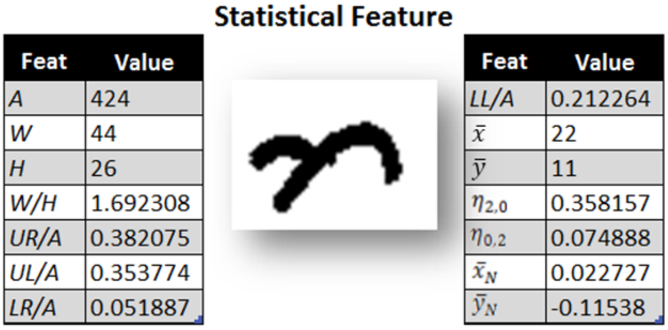
Normalized mass center $(\bar{x}_N, \bar{y}_N)$ is calculated by (4) dan (5).

$$\bar{x}_N = \frac{\bar{x} - (W-1)/2}{W/2} \quad (4)$$

$$\bar{y}_N = \frac{\bar{y} - (H-1)/2}{H/2} \,. \quad (5)$$

The width of the image is denoted by W, the height of the image is H and the ratio of the image width to the image height is W/H. From the four quadrants, there is a fractional form of black pixels with upper-right of area (UR/A), upper-left of area (UL/A), lower-right of area (LR/A) of area (LL/A).

For example, statistical feature of the letter of "ha" is shown in Fig. 12.



Statistical Feature

Feat	Value
A	424
W	44
H	26
W/H	1.692308
UR/A	0.382075
UL/A	0.353774
LR/A	0.051887

Feat	Value
LL/A	0.212264
$\bar{x}$	22
$\bar{y}$	11
$\eta_{2,0}$	0.358157
$\eta_{0,2}$	0.074888
$\bar{x}_N$	0.022727
$\bar{y}_N$	-0.11538

Fig. 12 Statistical feature of letter "ha".

2) Skeleton Features: In this category there are three features used, namely the number of branching points (BP), the number of end points (EP), as well as the number of points normal (NP) i.e. points other than BP and EP, on frame (skeleton) of the letter image. The illustrations of these three points are shown in Fig. 13.

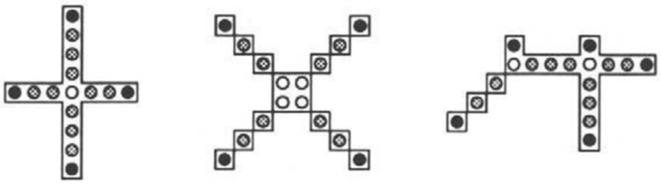


Fig. 13 Illustration for EP (black), BP (white), dan NP (grey).

3) Boundary Features: This feature is determined by the boundary of an image. There are four features that fall into this category. For example the outer coordinates of the image edge pixels are $(x(t), y(t))$, with $t = 1, 2, \ldots, m$.. The first feature taken is the number of outer pixels (borders) is m. Next Freeman's chain-code method is used to encode the outermost pixels $f(t) \in \{0, 1, \ldots, 7\}$, as shown in Fig. 14 [9].

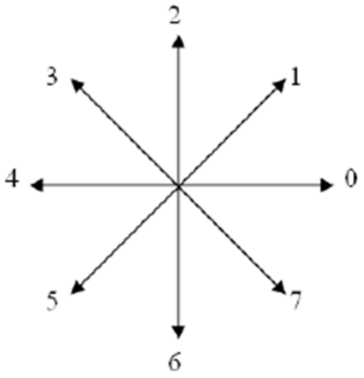


Fig. 14 The 8-connectivity in Freeman chain-code.

The perimeter length also applies as a feature, specified via (6).

$$T = \sum_{t=1}^{m} L(f(t)) \tag{6}$$

with

$$L(f(t)) = \begin{cases} 1 & \text{for even } f(t) \\ \sqrt{2} & \text{otherwise} \end{cases} \tag{7}$$

The perimeter to diagonal ratio is determined using the following formula:

$$\frac{T}{2D} = \frac{T/2}{\sqrt{W^2 + H^2}} \tag{8}$$

and finally the ratio of density (compactness) is calculated using the following formula.

$$\gamma = \frac{T^2}{4\pi A} \tag{9}$$

4) Directional Features: There are four features in this category. The directional feature (direction) is extracted through the chain-code of the handwriting boundary image. Only four of the eight directions are relevant for use, the remaining four other directions are as reflected from the first four. The four features of the direction D_d to be $d = 0, 1, 2, 3$ defined as follows.

$$D_d = \sum_{t} C_d(f(t)) \tag{10}$$

with

$$C_d(f(t)) = \begin{cases} 1 & \text{for } f(t) \text{ mod } 4 = d \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

In this paper, contrary with the previous research, the directional features are not divided into several regions [2].

5) Elliptic Fourier Descriptor Features: In this category, the ideas of Kuhl and Giardina are used [10]. The outer edge margin of the image is a closed contour used in determining EFD. The four descriptors of order *n* are defined as follows.

$$a_n = \frac{T}{2n^2\pi^2} \sum_{i=1}^{m} \frac{\Delta x_i}{\Delta t_i} [\cos \phi_i - \cos \phi_{i-1}] \tag{12}$$

$$b_n = \frac{T}{2n^2\pi^2} \sum_{i=1}^{m} \frac{\Delta x_i}{\Delta t_i} [\sin \phi_i - \sin \phi_{i-1}] \tag{13}$$

$$c_n = \frac{T}{2n^2\pi^2} \sum_{i=1}^{m} \frac{\Delta y_i}{\Delta t_i} [\cos \phi_i - \cos \phi_{i-1}] \tag{14}$$

$$d_n = \frac{T}{2n^2\pi^2} \sum_{i=1}^{m} \frac{\Delta y_i}{\Delta t_i} [\sin \phi_i - \sin \phi_{i-1}] \tag{15}$$

with

$$\phi_i = \frac{2n\pi t_i}{T} \tag{16}$$

$$\Delta x_i = x(i) - x(i-1) \quad \Delta y_i = y(i) - y(i-1) \tag{17}$$

$$\Delta t_i = \sqrt{\Delta x_i^2 + \Delta y_i^2} \tag{18}$$

$$t_i = \sum_{j=1}^{i} \Delta t_j \qquad T = t_m = \sum_{j=1}^{m} t_j \tag{19}$$

In this paper, 6-order EFD is selected as a feature, since in this order there has been a fairly good handwritten image reconstruction, as shown in Fig. 15 and Fig. 16.

An example of a 6-order EFD feature for the letter "ha" is shown in Fig. 17.

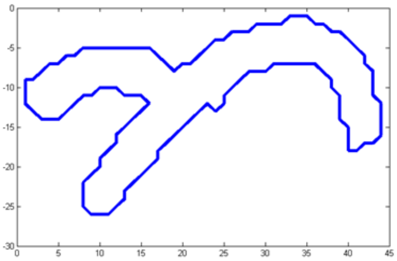


Fig. 15 Outer-contour of letter “ha”.

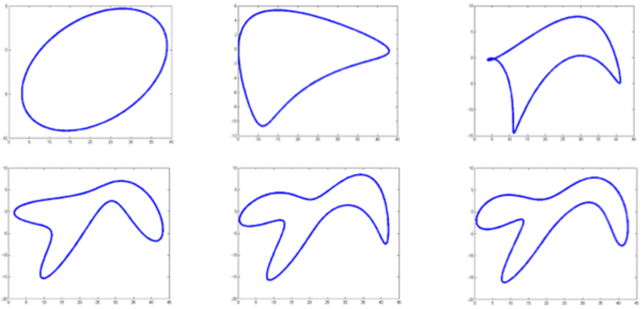


Fig. 16 Reconstruction of letter “ha” by EFD order 1-2-3 (up) and order 4-5-6 (down).

C. Classification Stage

In this paper, there are several classifiers used as a comparison, the method of Nearest Neighbors (NN), Naïve Bayes (NB), and Mahalanobis Distance (MD). These classifiers were chosen because of their fairly common use in pattern recognition [11], [12].

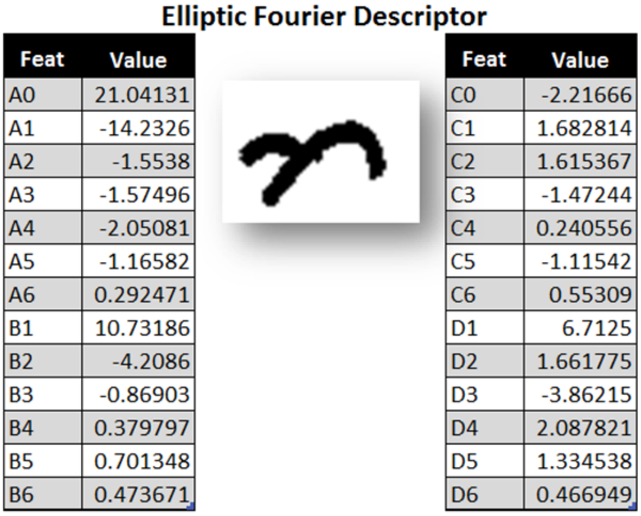
1) Nearest Neighbors (1-NN): This classifier determines the closest distance of the test feature to the sample feature used as a reference. The closest distance is calculated using Euclidean distance.

2) Naïve Bayes (NB): This classifier uses Bayes rules in search of a posteriori opportunities as follows:

$$P(\omega_i|x) = \frac{P(\omega_i)P(x|\omega_i)}{P(x)} \, . \tag{20}$$

Assuming each feature is independent, then the above formula becomes:

$$P(\omega_i|x) = \frac{P(\omega_i)P(x_1|\omega_i)P(x_2|\omega_i)\ldots P(x_n|\omega_i)}{P(x_1)P(x_2)\ldots P(x_n)} \tag{21}$$



Elliptic Fourier Descriptor

Feat	Value	Feat	Value
A0	21.04131	C0	-2.21666
A1	-14.2326	C1	1.682814
A2	-1.5538	C2	1.615367
A3	-1.57496	C3	-1.47244
A4	-2.05081	C4	0.240556
A5	-1.16582	C5	-1.11542
A6	0.292471	C6	0.55309
B1	10.73186	D1	6.7125
B2	-4.2086	D2	1.661775
B3	-0.86903	D3	-3.86215
B4	0.379797	D4	2.087821
B5	0.701348	D5	1.334538
B6	0.473671	D6	0.466949

Fig. 17 The EFD feature order-6 of letter “ha”.

The recognition process is performed by looking for the largest posteriori, so that (21) can be expressed as (22).

$$P(\omega_i|x) \propto P(\omega_i)\prod_{j=1}^{n} P(x_j|\omega_i) \tag{22}$$

The probability value $P(x_j|\omega_i)$ is estimated using the Parzel window with the Gaussian kernel function.

3) Mahalanobis Distance (MD): MD calculates the distance between the features tested with the features of each class using the following formula:

$$D_m^2 = (x-\mu)\Sigma^{-1}(x-\mu)^T \tag{23}$$

with μ is the mean of each feature in a class, Σ is the covariance matrix of each feature, x is the feature that Mahalanobis will look for.

IV. Results

Fig. 18, Fig. 19, and Fig. 20 show the recognition of each handwriting letter of Batak Toba script and the results from various feature extraction using three classifiers, namely MD, 1-NN, and NB.

In Fig. 18, it appears that handwriting recognition of the Batak Toba script using MD gives average recognition results based on feature extraction of EFD of 86.51%, Statistical 74.01%, Directional 57.07%, Boundary 54.93%, and Skeleton of 44 , 24%.

While the the handwriting recognition of the Batak Toba script using 1-NN, as in Fig. 19, it is obtained average recognition based on feature extraction of EFD: 90.13%, Directional 56.09%, 46,22% Boundary, Statistical 44.90%, and Skeleton 16.94%.

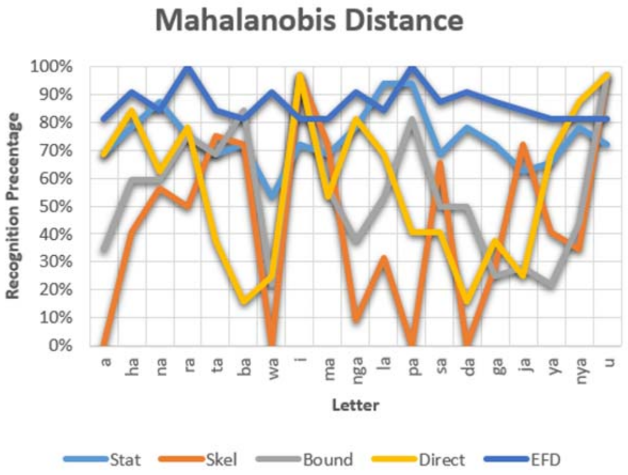


Fig. 18 Recognition by using MD Classifier.

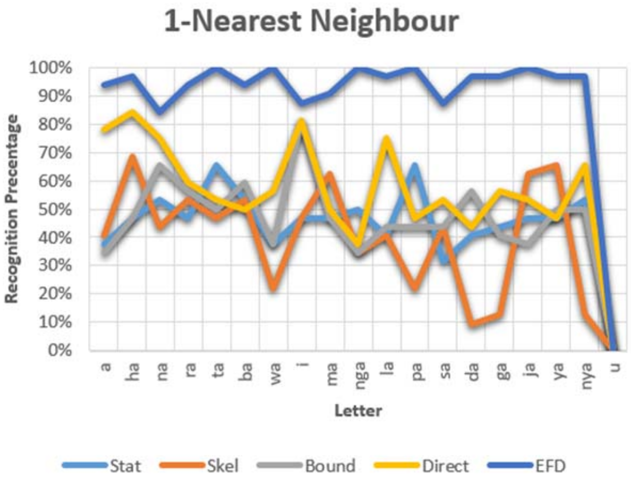


Fig. 19 Recognition by using 1-NN Classifier.

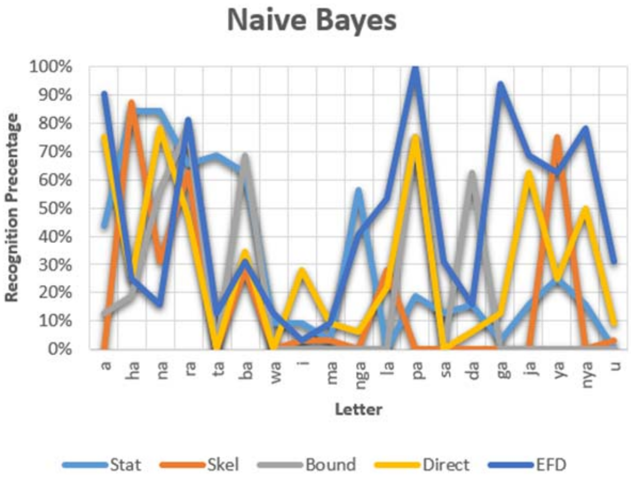


Fig. 20 Recognition by using NB Classifier.

Fig. 20 indicates that the handwriting recognition of the Batak Toba script using NB gives average recognition results based on EFD feature extraction: 45.07%, Statistical 31.25%, Directional 29.77%, Boundary 29.77%, and Skeleton is 16.94%.

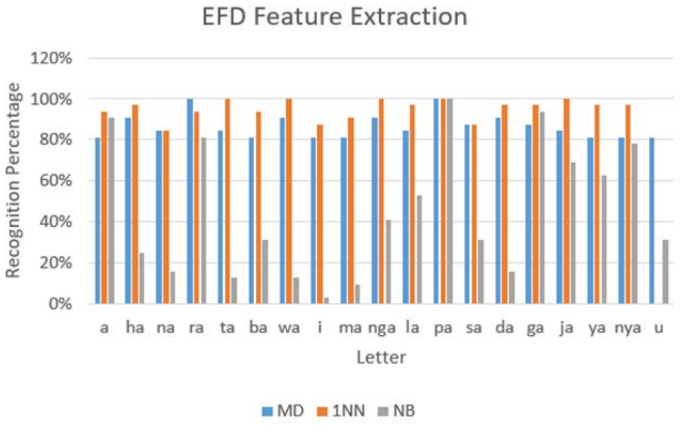


Fig. 21 Recognition comparison with EFD feature extraction.

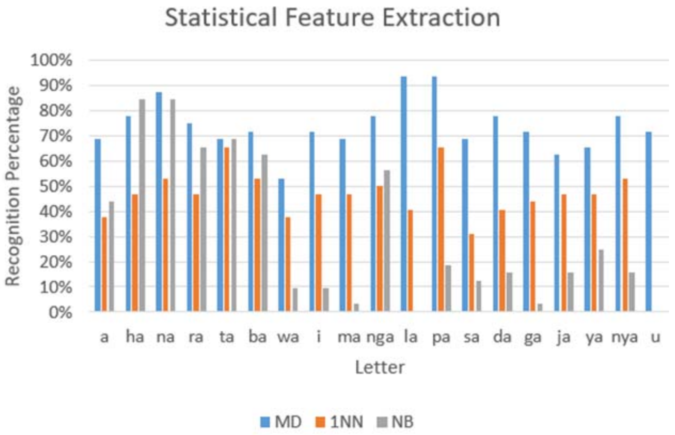


Fig. 22 Recognition comparison with statistical feature extraction.

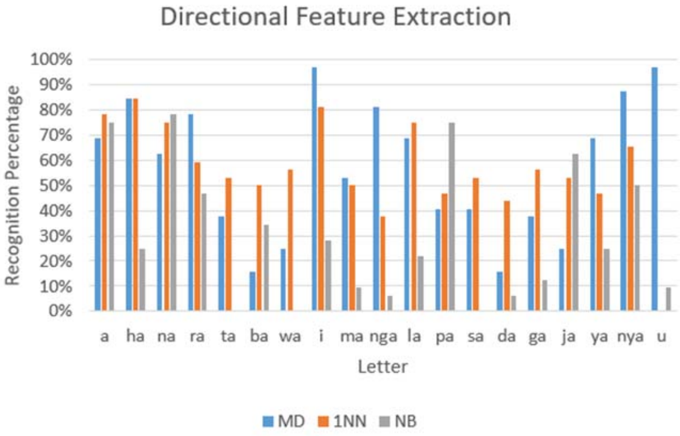


Fig. 23 Recognition comparison with Directional feature extraction.

Based on testing using three classifier of MD, 1-NN, and NB, it can be concluded that three feature extraction categories that produce the highest average recognition in sequence are *EFD*, *Statistical*, and *Directional*.

Furthermore, the comparison of the performance of the three classifiers based on the Batak Toba script handwriting recognition of each feature extraction is shown by Fig. 21 to Fig. 25.

Fig. 21 shows the results of the introduction by feature extraction of EFD using MD has the highest recognition percentage that is equal to 81.25% - 100%, while for NB is equal to 3.13% - 100%, and 1-NN is equal to 0% - 100%.

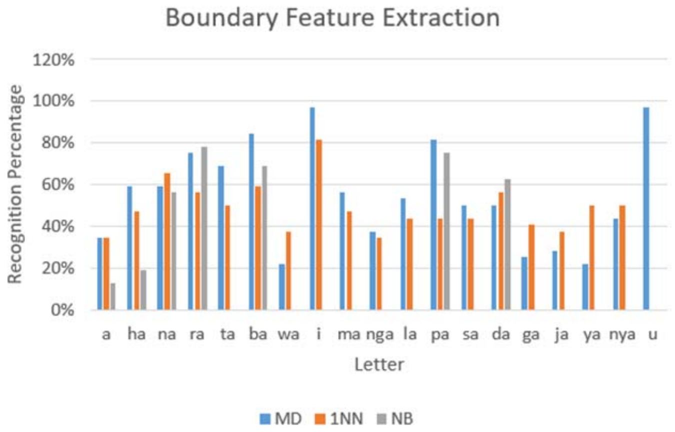


Fig. 24 Recognition comparison with Boundary feature extraction.

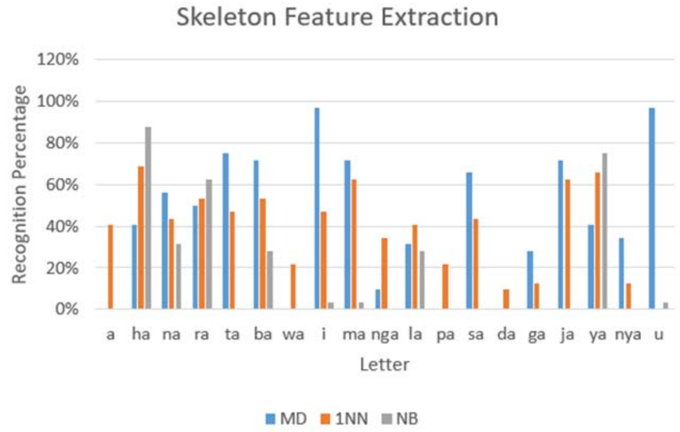


Fig. 25 Recognition comparison with Skeleton feature extraction.

Handwriting recognition with statistical feature extraction using MD, as in Fig. 22, has the highest recognition percentage of 53.13% - 93.75%, while for NB is 0% - 84.38%, and 1-NN is 0% - 65.63%.

Handwriting recognition results with Directional feature extraction are shown in Fig. 23. The highest recognition percentage is obtained by using MD, which is about 15.63% - 96.88%, while for 1-NN is equal to 0% - 84.38%, and NB is equal to 0% - 78.13%.

Fig. 24 shows the results of the recognition by feature extraction using Boundary MD of 21.88% - 96.88%, while for 1-NN is 0% - 81.25%, and NB is 0% - 78.13%.

In Fig. 25, shown the results of the recognition by Skeleton feature extraction. It can be seen that MD gives the highest recognition percentage, that is 0% - 96.88%, while for NB is 0% - 87.50%, and 1-NN is 0% - 68.75%.

From the comparison of the use of the three classifiers, it can be concluded that MD yields the highest percentage of the recognition, compared with NB and 1-NN.

V. Conclusion

This paper describes the process of the handwriting recognition of the Batak Toba script from data acquisition process, pre-processing stage, feature extraction, and classification. From the experiment, it can be concluded that the feature extraction category that give the highest recognition result (sequential) are EFD, Statistical, and Directional. In addition, from the three classifiers used, MD shows a better performance than the other two classifiers.

References

[1] U. Kozok, *Surat Batak: Sejarah Perkembangan Tulisan Batak, Berikut Pedoman Menulis Aksara Batak dan Cap Sisimangaraja XII* KPG (Kepustakaan Populer Gramedia) & EFEO, December 2015.

[2] G. A. Abandah, F.T. Jamour, E.A. Qaralleh, "Recognizing Handwritten Arabic Words using Grapheme Segmentation and Recurrent Neural Network", *International Journal on Document Analysis and Recognition (IJDAR)*, Springer-Verlag Berlin Heidelberg, 2014.

[3] D. Gosh, T. Dube, and A.P. Shivaprasad, "Script Recognition – A Review", *IEEE Trans. on PAMI*, 2009.

[4] (2016) SoBAT (Script of Batak Toba) Database website. [Online], http://aksarabatak.maranatha.edu, tanggal accessed 27 October 2016.

[5] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 1, 1979, pp. 62-66.

[6] Gonzalo Pajares, Jesus Manuel de la Cruz, "A Wavelet-based Image Fusion Tutorial," *Pattern Recognition*, March 2004.

[7] R. G. Casey dan E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation," *IEEE Trans. on PAMI* vol. 18, no 7, pp. 690-706, 1996.

[8] N. Theresia Br Pasaribu, M. J. Hasugian, "Noise Removal on Batak Toba Handwritten Script using Artificial Neural Network", *Proc. of 2016 3rd Int. Conf. on Information Tech., Computer, and Electrical Engineering (ICITACEE)*, pg. 373-76, Oct 19-21st, 2016, Semarang, Indonesia.

[9] H. Freeman, "On the Encoding of Arbitrary Geometric Configurations", *IRE Transactions on Electronic Computers. 10(2), 260–268*, 1961.

[10] F. Kuhl, C. Giardina, "Elliptic Fourier Features of a Closed Contour", *Computer Graphic and Image Processing 18(3), 236–258*, 1982.

[11] H. Peng, F. Long, C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", *IEEE Trans. on PAMI*, Vol 27, No 8, Agt 2005.

[12] G. Dougherty, *Pattern Recognition and Classification An Introduction*, Springer, New York, 2013.