

International Journal of Artificial Intelligence

ISSN 0974-0635

[HOME](#) [ABOUT](#) [LOGIN](#) [REGISTER](#) [SEARCH](#) [CURRENT](#) [ARCHIVES](#) [EDITORIAL BOARD](#) [AUTHOR INSTRUCTIONS](#)

SUBSCRIPTIONS

Home > Archives > 2020 Spring (March), Volume 18, Number 1

2020 Spring (March), Volume 18, Number 1

[Open Access](#) [Subscription or Fee Access](#)

Table of Contents

Part A: Regular Issue

A Hybrid Swarm Algorithm for Collective Construction of 3D Structures	PDF	1-18
<i>Henry Zapata, Niriasca Perozo, Wilfredo Angulo, Joyne Contreras</i>		
Multi-Layer Auto Resonance Network for Robotic Motion Control	PDF	19-44
<i>Vadanical Mathada Aparanji, Udayakumar Veerappa Wali, Ramalingappa Aparna</i>		
Conflict Monitoring Optimization Heuristic Inspired by Brain Fear and Conflict Systems	PDF	45-62
<i>Mojtaba Moattari, Mohammad Hassan Moradi</i>		
Whale Optimization Algorithm for Performance Improvement of Silicon-On-Insulator FinFETs	PDF	63-81
<i>Gurpurneet Kaur, Sandeep Singh Gill, Munish Rattan</i>		

Part B: Special Issue

Editorial: Special Issue on International Conference on Informatics, Robotics, Network, Control and Systems (IRONCONS)	PDF	82-85
<i>Endra Joelianto, Arjon Turnip, Augie Widyatratmo, Poltak Sihombing</i>		
Cost Analysis of Lemon Law Warranties for Used Equipments	PDF	86-96
<i>Hennie Husniah, R. Wangsaputra, Bermawi P. Iskandar</i>		
Design and Simulation of Traffic Light Control System at Two Intersections Using Max-Plus Model Predictive Control	PDF	97-116
<i>Endra Joelianto, Herman Y. Sutarto, Davindra Giovanni Airulla, Muhammad Zaky</i>		
Feature Extraction and Selection in Batak Toba Handwritten Text Recognition	PDF	117-134
<i>Novie Theresia Br Pasaribu, M. Jimmy Hasugian</i>		
Genetic Algorithms for Optimization of Multi-Level Product Distribution	PDF	135-147
<i>Asyrofa Rahmi, Wayan F. Mahmudy, M. Zoqi Sarwani</i>		
Management the Quality Control of Application the Adhesive on a Flat Material	PDF	148-162
<i>Peter Antony, Anna Antonyová, Endra Joelianto</i>		
Collision-Free Stable Polygon Formation of Multi-Agent Robots	PDF	163-176
<i>Azka M. Burohman, Endra Joelianto, Augie Widyatratmo</i>		
Lease Contracts with Servicing Strategy Model for Used Product Considering Crisp and Fuzzy Usage Rates	PDF	177-192
<i>Hennie Husniah, Asep K. Supriatna, Bermawi P. Iskandar</i>		
Hybrid Controller Design based Magneto-rheological Damper Lookup Table for Quarter Car Suspension	PDF	193-206
<i>Arjon Turnip, Jonny H. Panggabean</i>		
An Application of Modified Filter Algorithm Fetal Electrocardiogram Signals with Various Subjects	PDF	207-217
<i>Arjon Turnip, Andrian, Mardi Turnip, Abdi Dharma, Debora Paninsari, Tiarnida Nababan, Crismis Novalinda Ginting</i>		
Rehabilitation Procedure and Performance Measurement using Mechanical Rotary Impedance Actuator	PDF	218-236
<i>Augie Widyatratmo, Cahyoni Maharnani, Suprijanto</i>		
Combining Fuzzy Signature and Rough Sets Approach for Predicting the Minimum Passing Level of Competency Achievement	PDF	237-249
<i>Umi Laili Yuhana, Nurul Zainal Fanani, Eko Mulyanto Yuniarno, Siti Rochimah, Laszlo T. Koczys, Mauridhi Hery Purnomo</i>		

Disclaimer/Regarding indexing issue:

We have provided the online access of all issues and papers to the indexing agencies (as given on journal web site). **It's depend on indexing agencies when, how and what manner they can index or not.** Hence, we like to inform that on the basis of earlier indexing, we can't predict the today or future indexing policy of third party (i.e. indexing agencies) as they have right to discontinue any journal at any time without prior information to the journal. So, please neither sends any question nor expects any answer from us on the behalf of third party i.e. indexing agencies. Hence, we will not issue any certificate or letter for indexing issue. Our role is just to provide the online access to them. So we do properly this and one can visit indexing agencies website to get the authentic information.

SUBSCRIPTION

[Login to verify subscription](#)
[Give a gift subscription](#)

USER

Username
Password
 Remember me

NOTIFICATIONS

- [View](#)
- [Subscribe](#)

JOURNAL CONTENT

Search
Search Scope

Browse

- [By Issue](#)
- [By Author](#)
- [By Title](#)
- [Other Journals](#)

FONT SIZE

INFORMATION

- [For Readers](#)
- [For Authors](#)
- [For Librarians](#)

Feature Extraction and Selection in Batak Toba Handwritten Text Recognition

Novie Theresia Br Pasaribu ¹, and M. Jimmy Hasugian ^{2*}

Electrical Engineering Department, Maranatha Christian University, Indonesia
Email: ¹ novie.theresia@eng.maranatha.edu, ² jhasugian@maranatha.edu

ABSTRACT

Batak Toba is one of the tribes in Indonesia that has its own language and alphabet. However, not many Batak Toba people are familiar with the alphabet. In this paper, we use several feature extraction methods in the recognition process of Batak Toba handwritten text. For some features, between/within-class scatter matrix criterion is used to select the significant features. The k-NN classifier is used in the recognition step. The results show that elliptic Fourier descriptor is the most superior features that has recognition percentage greater than the other categories.

Keywords: Batak Toba text, Feature extraction, Elliptic Fourier descriptor, Feature selection, k-NN classifier.

Mathematics Subject Classification: 62H35, 68U10, 62H99, 68U99

Computing Classification System (1998): I.4.7, I.5.4

Computing Classification System (2012): Computing methodologies~Machine learning~Learning paradigms~Supervised learning; Mathematics of computing~Probability and statistics~Multivariate statistics

1. INTRODUCTION

Batak Toba is one of the tribes in Indonesia that has its own language and alphabet. Among the Batak Toba people, many of them are not familiar with the Batak Toba alphabet, (Kozok, 2015). This research is an attempt to conserve the culture of Batak Toba especially the alphabet through recognizing the handwritten. According to (Kozok, 2015) and (Ghosh, et al., 2009), Batak Toba alphabet can be said more likely a kind of an *abugida* in writing system, which is an alphabetic-like writing system. The Batak Toba alphabet consists of 19 *ina ni surat* (primary letter) and 6 *anak ni surat* (appear as a diacritic to the primary letter that changes the pronunciation) without any distinction between upper-case and lower-case letters (Pasaribu & Hasugian, 2016).

One main issue that occurs in pattern recognition system is in the determination of the data representation method (Jain, et al., 2000), so that the recognition rate can be improved. This is known as the determination of features extraction and features selection. The Batak Toba alphabets can be categorized as Brahmic family script similar with Assamese character (Sarma, 2009) and Arabic letters (Abandah, et al., 2014). The dominant characteristic of Batak Toba letters is the presence of stroke and curve which also appear those two letters. This similarity is the reason we adopted some of the features extraction that are used in (Abandah, et al., 2014). However, the Batak Toba alphabet

is a non-cursive style alphabet, so not all features extraction that are applied in Arabic letters can be adopted in Batak Toba's handwritten circumstances.

According to our knowledge, there is only one publication that has been reported about Batak Toba character recognition (Panggabean & Rønningen, 2009). In their study, the printed Batak Toba font-type was used as the character to be recognized. These Batak Toba font types were downloaded at (Kozok, 2008). Rather than using a font type, whereas we use Batak Toba handwritten text, the problems raised are more complex as the result of the variation in human handwriting style.

In this study, we propose the Batak Toba handwritten text recognition system from the pre-processing step, extracting the significant features, feature selection process, and finally the classification process by using k-NN approach as the classifier.

2. METHODOLOGY

Batak Toba alphabet is usually called as *si sia-sia* or *surat sampulu sia* because the number of the letter (*ina ni surat*) in Batak Toba are nineteen (sampulu sia means nineteen) (Kozok, 2015). The nineteen letters are presented in Fig. 1.

The whole proposed system is depicted in a block diagram (see Fig. 2). In this study, the handwriting from hundreds of high school students in Saposurung Balige, North Sumatera, Indonesia (see Fig. 3) who are familiar with the Batak Toba characters, were utilized as our dataset. The students were asked to write the Batak Toba alphabet (Fig. 1) in a piece of A4 paper. All of these handwriting databases is available at our website (Pasaribu & Hasugian, 2016).

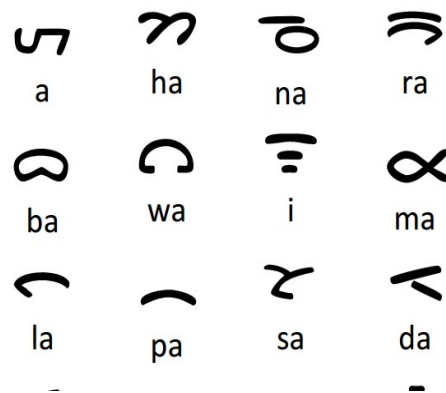


Figure 1. The modern version of *ina ni surat*

Then the handwritten text papers were scanned at 300 dpi resolution and segmented afterwards by using the horizontal projection method for line segmentation and the vertical projection method for character segmentation (Casey & Lecolinet, 1996) (Cheriet, et al., 2007).

The horizontal and vertical projection methods respectively are calculated by using these following equations (Sonka, et al., 2015):

$$H[i] = \sum_{j=1}^n B[i, j] \quad (1)$$

$$V[j] = \sum_{i=1}^m B[i, j] \quad (2)$$

where $B[i, j]$ represents the pixel intensities of the handwritten text image.

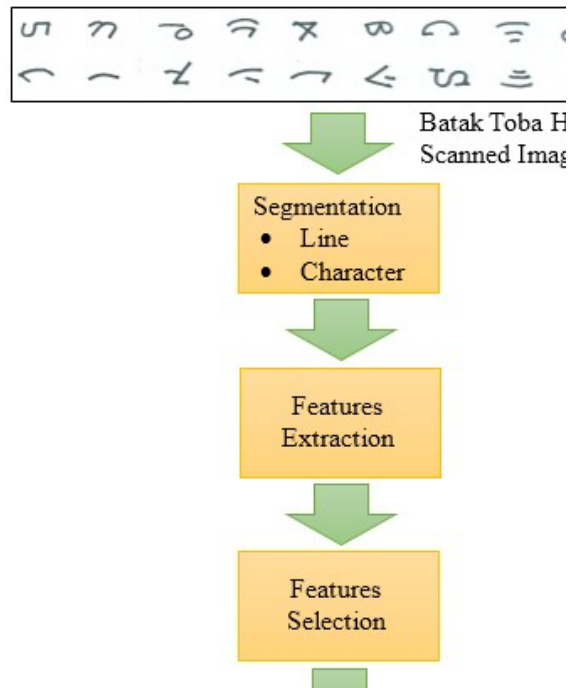


Figure 2. Block diagram of the Batak Toba handwritten recognition system



Figure 3. High school students in Soposurung wrote Batak Toba alphabet

The segmentation process is illustrated in Fig. 4.

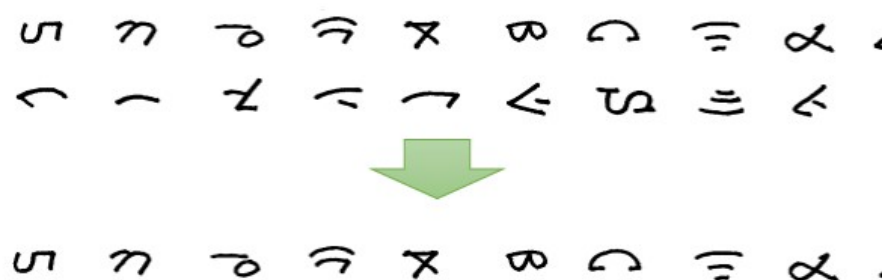


Figure 4. The process for line and character segmentation

The segmented letters then organized into letters dataset to train and test the system. The example of letters collection is presented in Fig. 5.

Alphabet	Respondent									
	1	2	3	4	5	6	7	8	9	10
a										
ha										
na										
ra										
ta										
ba										
wa										
i										
ma										
nga										
la										
pa										
sa										
da										
ga										
ja										
ya										
nya										
u										

Figure 5. The sample of Batak Toba alphabet from ten respondents' handwriting

The next step is then to extract some common features and to select the significant features before the classification process. A thorough discussion of these topics is described in the following sections.

3. FEATURES EXTRACTION

Features extraction used in this paper is divided into five categories i.e. statistical, skeleton, boundary, directional, and elliptic Fourier descriptors (EFD), (Yampolskiy, 2004), (Abandah, et al., 2014).

3.1 Statistical features

The statistical feature is the most common feature used in the scheme of the character recognition system. In this category, 14 features are extracted. Area A is determined through this equation:

$$A = \sum_x \sum_y B(x, y) \quad (3)$$

with $B(x, y)$ is the image intensity of binary image.

The *width* (W) and *height* (H) of an object and its *ratio* (W/H) are used as features in this category. Then the division of foreground pixels in each quadrant i.e. *upper-right of area* (UR/A), *upper-left of area* (UL/A), *lower-right of area* (LR/A), *lower-left of area* (LL/A) are also counted in this category.

The *center of mass* of the object in each coordinate (\bar{x}, \bar{y}) is exploited to calculate the *normalized central moments* $\eta_{2,0}$ and $\eta_{0,2}$ by using the following formulas

$$\eta_{u,v} = \frac{1}{A^k} \sum_x \sum_y (x - \bar{x})^u (y - \bar{y})^v \cdot B(x, y) \quad (4)$$

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

The *normalized center of mass* (\bar{x}_N, \bar{y}_N) are calculated by:

$$\bar{x}_N = \frac{\bar{x} - (W - 1)/2}{W/2} \quad (5)$$

$$\bar{y}_N = \frac{\bar{y} - (H - 1)/2}{H/2} \quad (6)$$

As an example, the statistical feature of letter “ha” is depicted in Fig. 5.

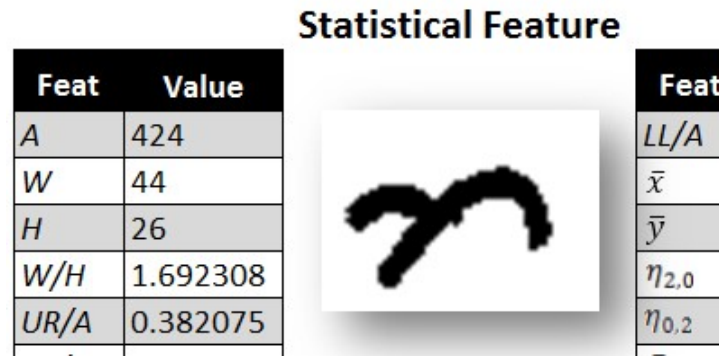


Figure 6. The statistical feature of letter “ha”

3.2 Skeleton features

In this category, three features are excerpted from the object’s skeleton (Kong & Rosenfeld, 1996). *Branch points* (BP) is the number of pixels at the branch of object’s skeleton, and *end points* (EP) is the number of end points. In addition, *normal points* (NP) which are points other than BP and EP, are taken as features in this category. The position of BP, EP, and NP are illustrated in Fig. 7.

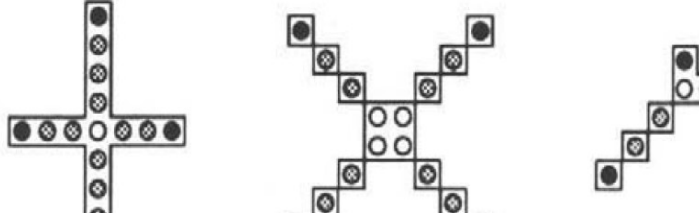


Figure 7. Illustration of EP (black circle), BP (white circle), and NP (gray circle)

3.3 Boundary features

From the object's boundary, four features are taken into consideration. The coordinate of the outer boundary pixel is $(x(t), y(t))$ for $t = 1, 2, \dots, m$. The *total pixel at the boundary* is m . Freeman chain-code (Freeman, 1961) is deployed to encode the pixel at the boundary. The orientation of every pixel to the adjacent pixel is put in a chain with codes $f(t) \in \{0, 1, \dots, 7\}$ as displayed in Fig. 8.

The *length of perimeter* T is determined by the following formula:

$$T = \sum_{t=1}^m L(f(t)) \quad (7)$$

where

$$L(f(t)) = \begin{cases} 1 & \text{for } f(t) \text{ even} \\ \sqrt{2} & \text{for } f(t) \text{ odd} \end{cases} \quad (8)$$

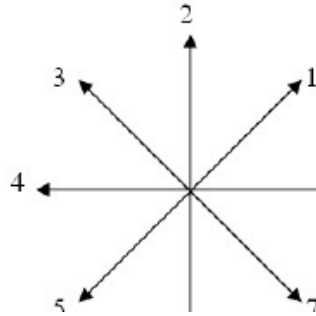


Figure 8. The 8-connectivity in Freeman chain-code

The *perimeter to diagonal ratio* is also considered as the feature in this category through the formula:

$$\frac{T}{2D} = \frac{T/2}{\sqrt{W^2 + H^2}} \quad (9)$$

And finally, the *compactness ratio* which is independent of the linear transformation:

$$\gamma = \frac{T^2}{4\pi A} \quad (10)$$

3.4 Directional features

In this category, the features are also taken from the object's boundaries chain codes, but only with four directions. The other four directions are assumed to be the reflection of the first four. The directional features D_d , $d = 0,1,2,3$ are defined as:

$$D_d = \sum_t C_d(f(t)) \quad (11)$$

where,

$$C_d(f(t)) = \begin{cases} 1 & \text{for } f(t) \bmod 4 = d \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

Different to (Abandah, et al., 2014), in this study, the object was not separated into any regions because Batak Toba character (letter) has no upper, middle, and lower zone as in Arabian letter.

3.5 Elliptic Fourier descriptors

In this category, the idea from (Kuhl & Giardina, 1982) is applied. This feature outperformed other methods based on closed contours (Trier, et al., 1996). The external boundary of an object is piecewise linear closed contour and is utilized in obtaining the elliptic Fourier descriptors (EFD). The four descriptors of order n are defined by

$$a_n = \frac{T}{2n^2\pi^2} \sum_{i=1}^m \frac{\Delta x_i}{\Delta t_i} [\cos \phi_i - \cos \phi_{i-1}] \quad (13)$$

$$b_n = \frac{T}{2n^2\pi^2} \sum_{i=1}^m \frac{\Delta x_i}{\Delta t_i} [\sin \phi_i - \sin \phi_{i-1}] \quad (14)$$

$$c_n = \frac{T}{2n^2\pi^2} \sum_{i=1}^m \frac{\Delta y_i}{\Delta t_i} [\cos \phi_i - \cos \phi_{i-1}] \quad (15)$$

$$d_n = \frac{T}{2n^2\pi^2} \sum_{i=1}^m \frac{\Delta y_i}{\Delta t_i} [\sin \phi_i - \sin \phi_{i-1}] \quad (16)$$

where

$$\phi_i = \frac{2n\pi x_i}{T} \quad \Delta x_i = x(i) - x(i-1) \quad \Delta y_i = y(i) - y(i-1)$$

$$\Delta t_i = \sqrt{\Delta x_i^2 + \Delta y_i^2} \quad t_i = \sum_{j=1}^i \Delta t_j \quad T = t_m = \sum_{j=1}^m t_j$$

We adapted the algorithm in (Bose, 2000) to calculate every EFD coefficient. In this study, the sixth order is chosen for EFD features, because it results best character reconstruction yet not expanding the number of coefficients (See Fig. 9 and Fig. 10).

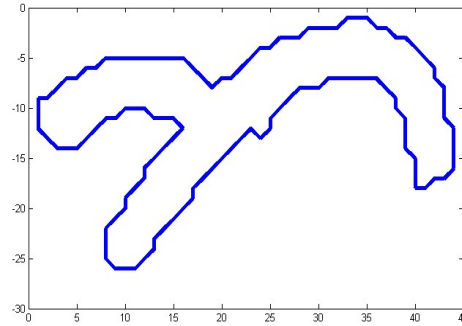


Figure 9. Outer-boundary contour of letter “ha”

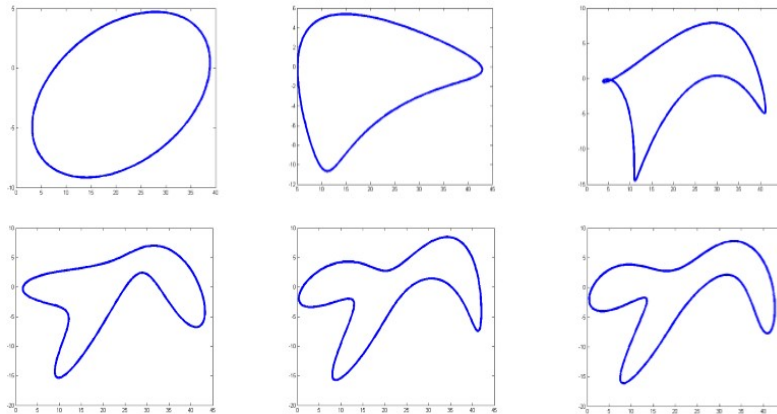



Figure 10. Reconstruction of Fourier coefficient. Order 1-2-3 (above) and order 4-5-6 (below) of the letter “ha”

By choosing the 6th order of EFD, the total coefficients are 26 on hand. The EFD features of letter “ha” is presented in Fig. 11 as an example.

Elliptic Fourier Descriptor

Feat	Value
A0	21.04131
A1	-14.2326
A2	-1.5538
A3	-1.57496
A4	-2.05081
A5	-1.16582
A6	0.292471
B1	10.73186
B2	-4.2086
B3	-0.86903



Feat
C0
C1
C2
C3
C4
C5
C6
D1
D2
D3

Figure 11. The 26 of the EFD feature of letter “ha”

4. FEATURE SELECTION AND CLASSIFIER

In many applications, some should deal with plenty of features. Therefore, eradicating irrelevant or even redundant features is an eminent step in pattern classification scheme. This process known as feature selection. Fundamentally, the goal is to aggregate the discriminative features. For this purpose, one should attain the high correlation between feature and target (relevancy) and at the same time avoid high correlation between feature-to-feature (redundancy) (Barchinezhad & Eftekhari, 2014).

Ideally, discriminative features should have less intra-class variability and more variability in inter-class as depicted in Fig. 12 (Dougherty, 2013). In this paper, the within-class scatter matrix S_w and between-class scatter matrix S_b (Johnson & Wichern, 2007) are exploited as a tool for selecting which feature in every category will improve the recognition rate.

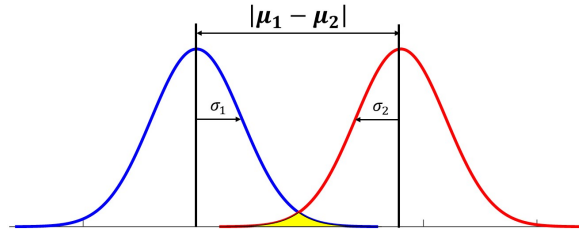


Figure 12. Illustration of intra/inter class variations

Suppose there are C classes and let μ_i be the mean vector of class $i=1,2,\dots,C$. Let N_i be the number of samples within class i . Then $N = \sum_{i=1}^C N_i$ be the total number of the sample.

Within-class scatter matrix is calculated by

$$S_w = \sum_{i=1}^C \sum_{j=1}^{N_i} (x_j - \mu_i)(x_j - \mu_i)^T \quad (17)$$

and between-class scatter matrix

$$S_b = \sum_{i=1}^C (\mu_i - \mu)(\mu_i - \mu)^T \quad (18)$$

where

$$\mu = \frac{1}{C} \sum_{i=1}^C \mu_i \quad (\text{mean of entire data sets}) \quad (19)$$

The objective in this step is to find the large between-class scatter matrix and the small within-class scatter matrix, through the ratio of trace matrix of S_b and trace matrix of S_w as described by Eq. (20).

$$J = \frac{\text{tr}(S_b)}{\text{tr}(S_w)} \quad (20)$$

where $\text{tr}(\cdot)$ is a trace of matrix (Anton & Rorres, 2014).

The classification stage is the final stage in inspecting the trait of each feature category. There are numbers of classification method have been developed. Some classic widely used classifiers are k -nearest neighbor (Barchinezhad & Eftekhari, 2014), linear discriminant analysis (Duda, et al., 2001), and support vector machine (Mu, et al., 2017). However, by the reason of simplicity, the k -NN classifier is chosen in this recognition process (Cover & Hart, 1967); (Zheng, et al., 2004). The k -NN algorithm begins at the test point \mathbf{x} then expands a zone until it encircle k training samples and then labels it according to a majority vote of these samples (Dougherty, 2013); (Duda, et al., 2001) as illustrated in Fig. 13.

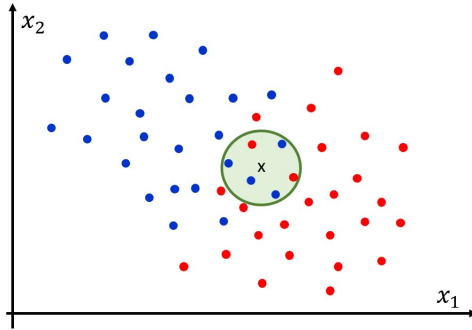


Figure 13. Illustration of k -NN process

Euclidean distance d_ε is employed as the proximity measure of the test point t_i with the samples s_i as described in Eq. (21) (Theodoridis & Koutroumbas, 2009)

$$\begin{aligned} d_\varepsilon &= \|t_i - s_i\| \\ &= \sqrt{(t_1 - s_1)^2 + \dots + (t_n - s_n)^2} \end{aligned} \quad (21)$$

5. EXPERIMENTAL SETUP AND RESULTS

In the first step, all categorical features are extracted from each segmented-character (letter), then applying the k -NN classifier (with 1 nearest neighbor) to recognize each letter in the Batak Toba alphabet. When the result is quite high, we assumed that the feature is already discriminative so that each letter distinguishable. Then the feature category which has a lower recognition rate undergoes a selection process. Finally, the results will be compared with the first step.

Three metrics are utilized to evaluate the results: the recognition rate in term of sensitivity (S_e) as the value to measure how good the system recognizes all the targeted letter correctly, specificity (S_p) to assess the ability of the system not to misclassify the particular letter, and precision (P_r) as a degree to estimate from all the letter that classified as a particular letter that certainly is from that letter. All evaluation metrics are determined in terms of TP (true positive), FN (false negative), TN (true negative), and FP (false positive) as follows:

$$S_e = \frac{TP}{TP + FN} \quad S_p = \frac{TN}{TN + FP} \quad P_r = \frac{TP}{TP + FP} \quad (22)$$

The recognition rate or sensitivity when operating the statistical features set is depicted in Fig. 14.

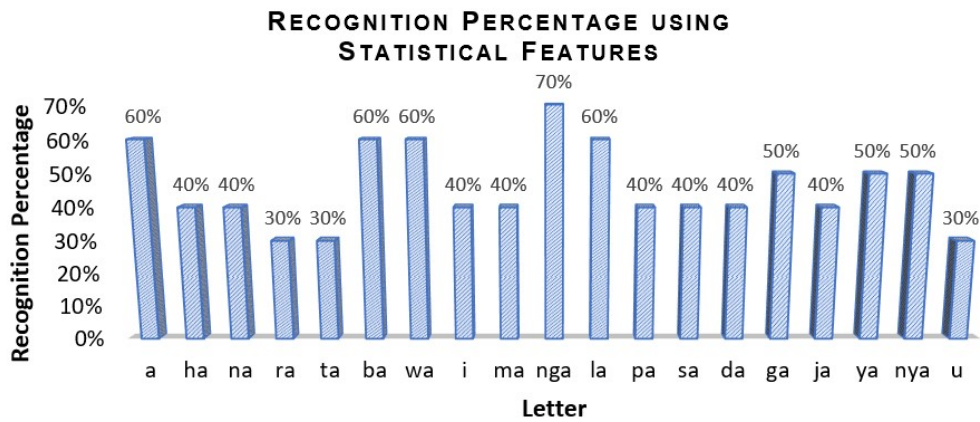


Figure 14. Recognition rate of statistical features set

The recognition rate range is between 30% and 70%, with the major rate 40% (8 out of 19 letters or 42% of letters). It can be said that the recognition by employing this feature is mediocre.

The recognition rate when using the skeleton features set is presented in Fig. 15.

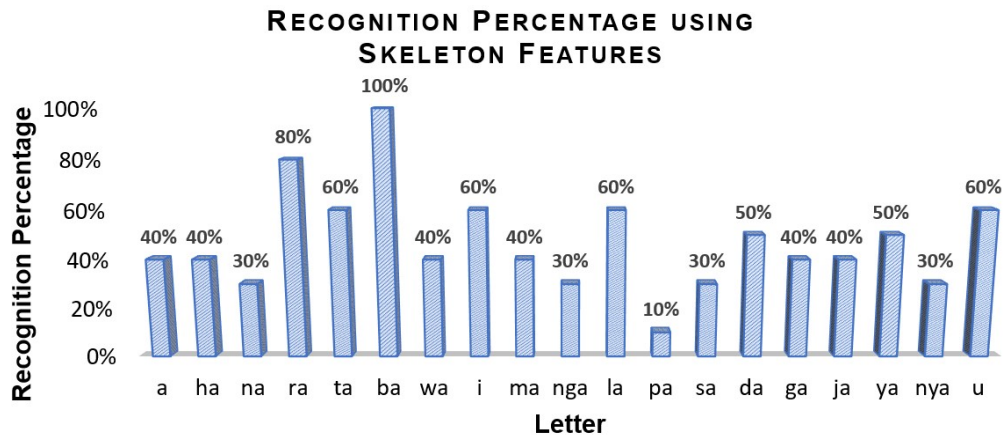


Figure 15. Skeleton features set recognition rate

From the above figure, the recognition rate is between 10% - 100%, and fluctuates around 40%. This variation of recognition rate expresses the instability of the system when taking the skeleton features.

The boundary features set recognition rate is displayed in Fig. 16.

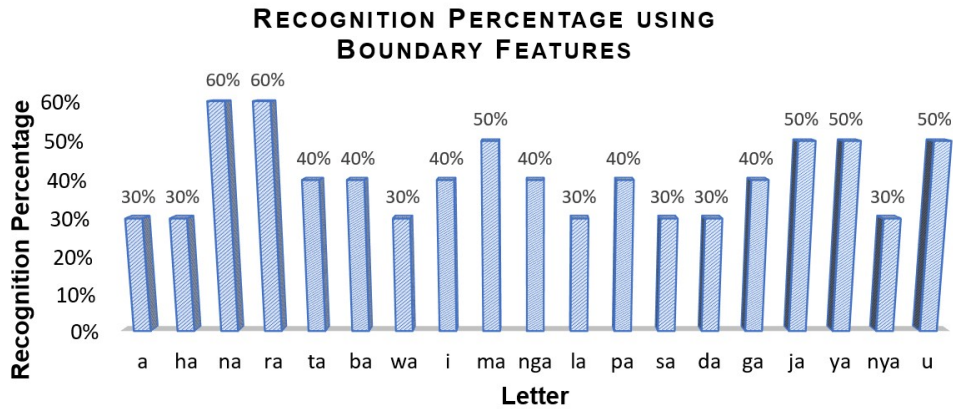


Figure 16. Recognition rate of boundary features set

It is shown that the recognition rate is between 30% and 60%. However, the majority is above 40%.

Fig. 17 presents the recognition rate when taking the directional features set as the input. The recognition is between 30% and 80%. On average, the recognition rate is 53% and relatively constant and dominated by 50%.

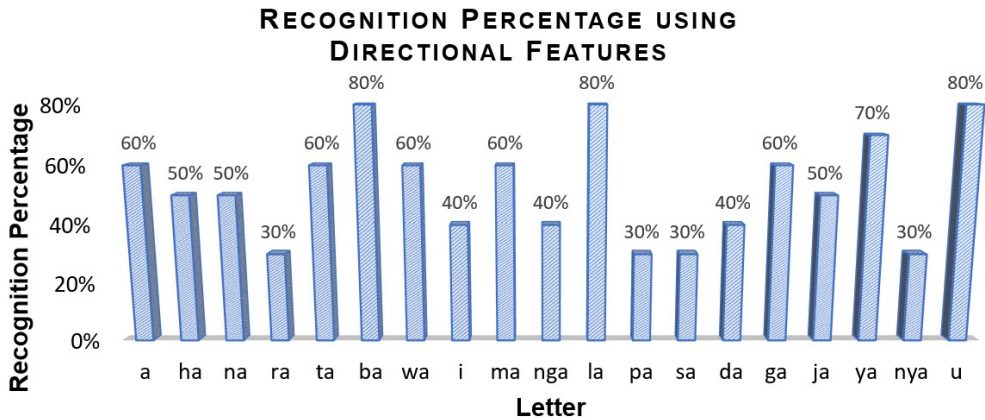


Figure 17. Directional features set recognition rate

Meanwhile, the recognition rate when using EFD features set is given in Fig. 18. The results show that the recognition rate is between 50% and 100%. And surprisingly the results are dominated by 90% and 100%.

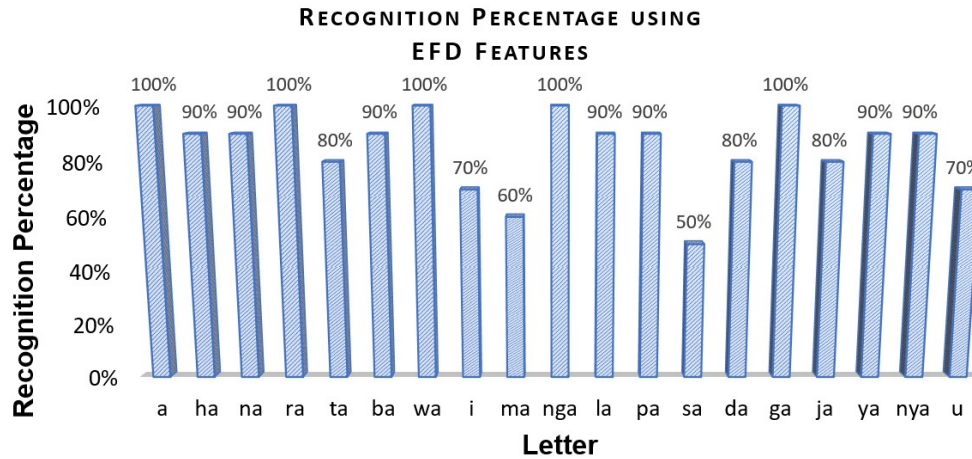


Figure 18. Recognition rate of EFD features set

In general, the comparison of the recognition rate in term of sensitivity of all features categories is depicted in Fig. 19.

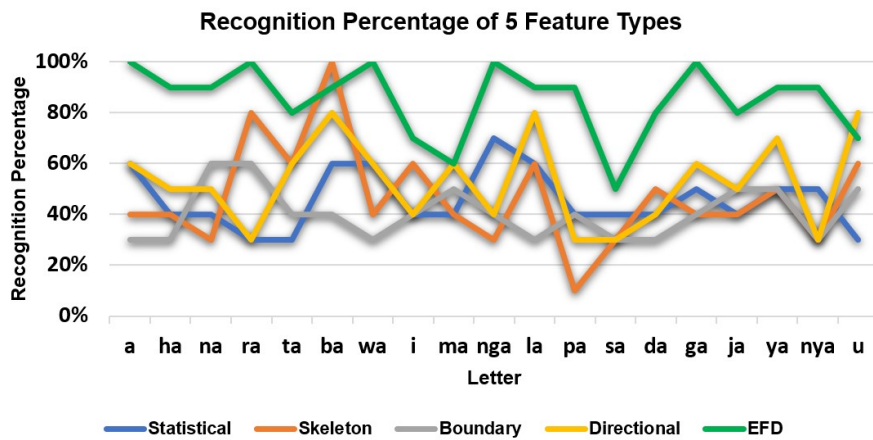


Figure 19. Comparison of recognition rate of all feature categories

The entire results for other two evaluation metrics are reported in Table 1. Fig. 19 and Table 1 shows that the EFD features has superior recognition rate as well as the specificity and precision compared to other feature categories. There are five letters (about 26 %) that can be recognized perfectly 100%, about 37 % that has 90 % recognition percentage, and the rest have fluctuated recognition from 50% until 80% accuracy. However, in the most of the letter's recognition, EFD feature set outperforms the others.

Table 1. The other metrics for all feature categories

Letter	Stat		Skel		Bound		Direct		EFD	
	Sp	Pr	Sp	Pr	Sp	Pr	Sp	Pr	Sp	Pr
a	0.97	0.50	0.94	0.27	0.95	0.23	0.97	0.55	0.99	0.91
ha	0.96	0.36	0.93	0.22	0.97	0.33	0.99	0.83	1.00	1.00
na	0.97	0.44	0.99	0.60	0.99	0.86	0.99	0.71	0.99	0.90
ra	0.96	0.27	0.97	0.62	0.95	0.40	0.98	0.50	0.99	0.91
ta	0.98	0.50	0.98	0.60	0.97	0.40	0.99	0.86	0.99	0.89
ba	0.98	0.60	0.95	0.53	0.97	0.40	0.95	0.44	0.99	0.90
wa	0.97	0.50	0.97	0.44	0.95	0.25	0.96	0.43	1.00	1.00
i	0.96	0.36	0.99	0.86	0.99	0.80	0.98	0.57	0.98	0.64
ma	0.98	0.50	0.97	0.40	0.96	0.42	0.97	0.50	0.99	0.75
nga	0.98	0.70	0.98	0.50	0.95	0.31	0.95	0.31	0.99	0.83
la	0.98	0.67	0.99	0.86	0.93	0.19	0.97	0.62	0.98	0.69
pa	0.99	0.80	0.98	0.25	0.99	0.67	0.99	0.60	0.99	0.90
sa	0.98	0.57	0.94	0.21	0.94	0.20	0.97	0.33	1.00	1.00
da	0.98	0.57	0.95	0.33	0.99	0.75	0.98	0.57	0.98	0.67
ga	0.97	0.50	0.96	0.33	0.98	0.50	0.98	0.60	0.99	0.91
ja	0.97	0.44	0.99	0.80	0.96	0.42	0.97	0.50	1.00	1.00
ya	0.97	0.45	0.98	0.56	0.99	0.83	0.98	0.70	1.00	1.00
nya	0.92	0.24	0.98	0.50	0.98	0.43	0.96	0.30	0.99	0.90
u	0.96	0.27	1.00	1.00	0.97	0.50	0.96	0.53	0.98	0.64

Since only the results based-on EFD features category has high recognition percentage, therefore the other features are taken selectively by using the feature selection criteria as described in the Eq. (20) to improve the recognition rate. The threshold for J is 5.00. It means, features that has J value less than 5.00 is removed. Then, the recognition process is conducted again through the new features-set.

The comparisons of recognition percentage before and after the feature selection process are displayed in the following figures.

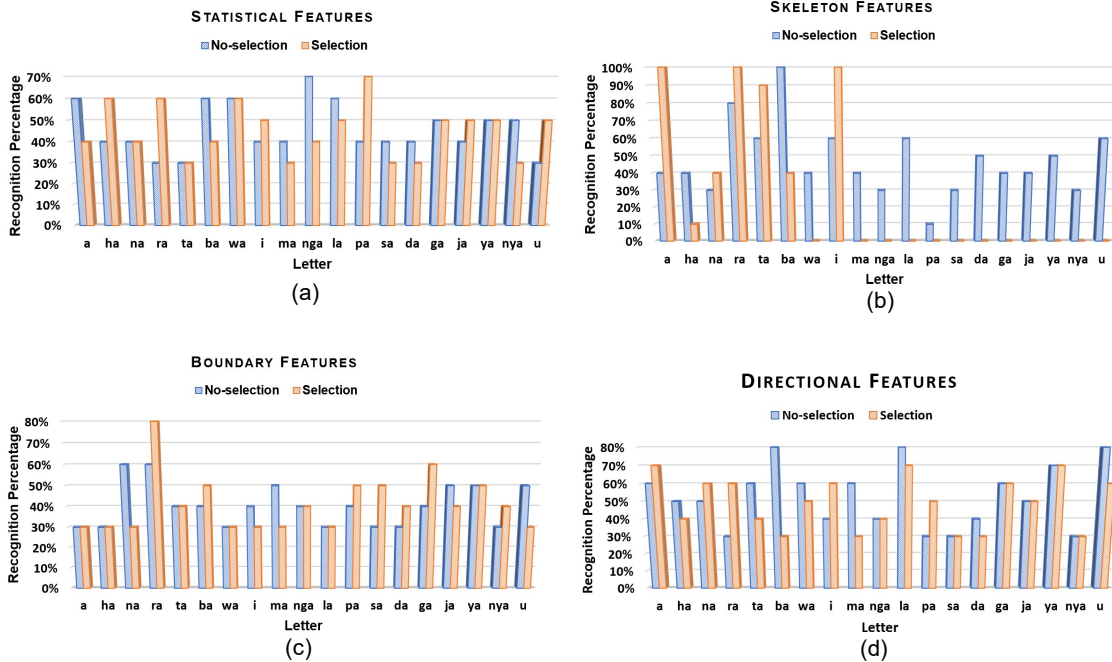


Figure 20. Recognition rate comparison before and after selection process, (a) statistical, (b) skeleton, (c) boundary, and (d) directional features

In statistical category (Fig. 20 a), there are only 6 out of 19 letters (32%) have recognition improvement, i.e. letters “ha”, “ra”, “i”, “pa”, “ja”, and “u”. The range of recognition rate is around 30% to 70%. However, the feature selection process gave poor impact to the recognition of letter “a”, “ba”, “ma”, “nga”, “la”, “sa”, “da”, and “nya” (8 letters). Similar with statistical, skeleton category (Fig 20 b) also has 5 out of 19 (26%) letters that show improvement in recognition percentage, i.e. letter “a”, “na”, “ra”, “ta”, and “i”. However, most of the recognition percentages are decrease sharply due to the selection process.

In boundary features category, it can be seen in Fig. 20 c, that there are 7 out of 19 letters (37%) show increasing of recognition rate significantly, i.e. letter “ra”, “ba”, “pa”, “sa”, “da”, “ga”, and “nya”. The range of recognition rate is improved from 30%-60% to 30%-80%. Only 5 letters (26%) show slight decline in the recognition rate, and the rest of the letters (37%) still at the same recognition percentage. Meanwhile in directional features category (Fig. 20 d) demonstrates the same numbers of improvement as the skeleton features category which is about 26 %, i.e. letter “a”, “na”, “ra”, “i”, and “pa”. And 32% (6 letters) show the same recognition rate before and after selection process, i.e. letter “nga”, “sa”, “ga”, “ja”, “ya”, and “nya”. Unfortunately, there are 8 letters (42%) have reduction in recognition percentage.

Specificity and precision comparison are reported in Fig. 21 and Fig. 22 respectively. In general, the changes of these metrics are quite similar with recognition rate, but the specificity of the skeleton features is constantly above 0.94, except for letter “a”.

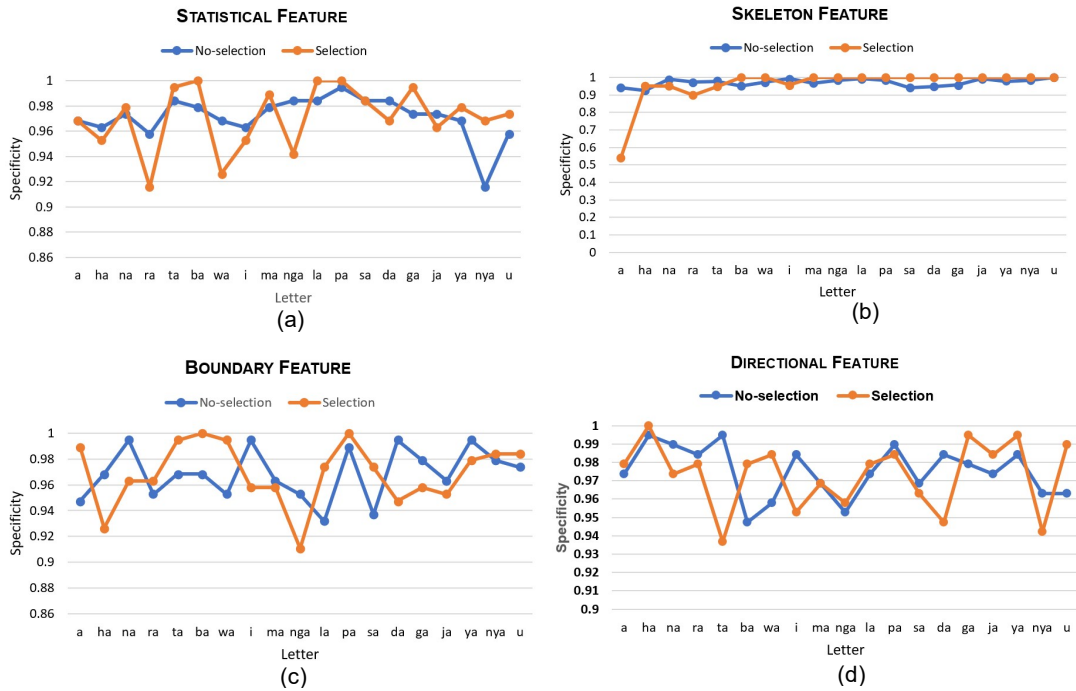


Figure 21. Comparison of specificity before and after selection, (a) statistical, (b) skeleton, (c) boundary, and (d) directional features

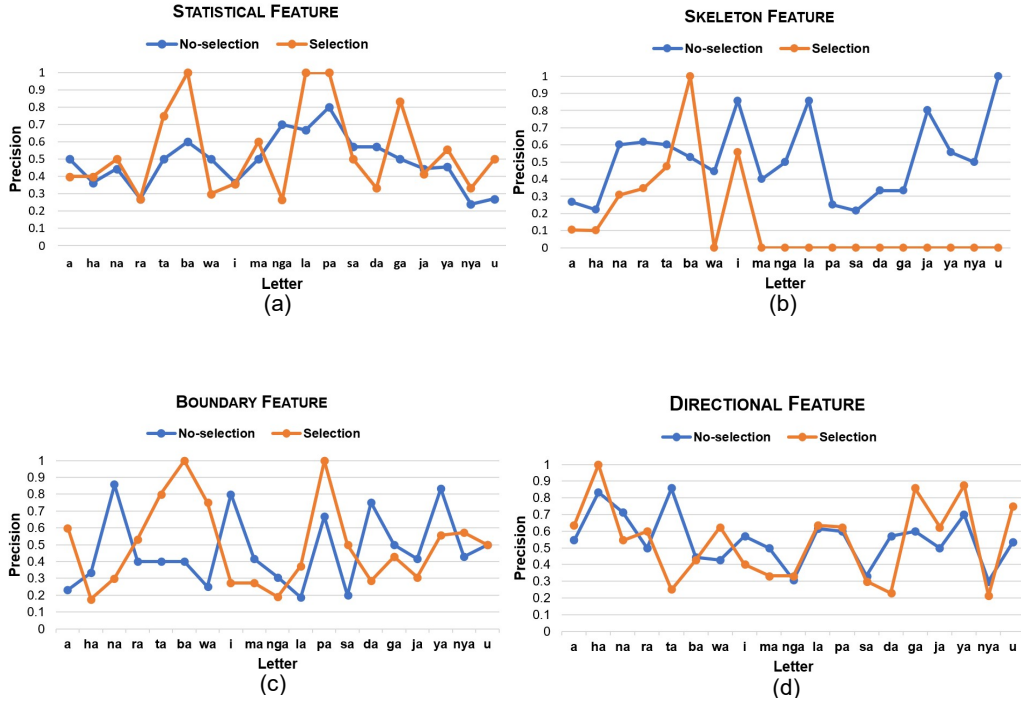


Figure 22. Precision comparison before and after selection, (a) statistical, (b) skeleton, (c) boundary, and (d) directional features

6. CONCLUSION

From the experiments, it can be inferred that the EFD features type is the most significant features in Batak Toba handwritten text recognition compared to statistical, skeleton, boundary, and directional features sets. At the average, it is more than 89% of Batak Toba handwritten recognition when using EFD as the feature extraction outperforms the other feature categories; which is 5 out of 19 letters can be recognized 100 % perfectly, i.e. letter “a”, “ra”, “wa”, “nga”, and “ga”. On the other hand, the feature selection procedures using the intra-class and inter-class scatter matrix criterion can only impact the statistical, boundary, and directional features set. The skeleton features show a sharp decrease after taking the feature selection procedures. These results bring suggestion to deploy another method for selection process. Finally, it can be concluded, despite of the classifier used in this process, EFD features extraction is the best features for Batak Toba handwritten text recognition.

7. REFERENCES

- Abandah, G. A., Jamour, F. T. & Qaralleh, E. A., 2014, Recognizing handwritten Arabic words using grapheme segmentation and recurrent neural networks. *International Journal on Document Analysis and Recognition*.
- Anton, H. & Rorres, C., 2014, *Elementary linear algebra: applications version*. 11th Ed. Wiley, USA.
- Barchinezhad, S. & Eftekhari, M., 2014, A new fuzzy and correlation based feature selection method for multiclass problems. *International Journal of Artificial Intelligence*, **12**(2), 24-41.
- Bose, P., 2000, *The encoding and fourier descriptors of arbitrary curves in 3-dimensional space*. Master's Thesis, University of Florida,
- Casey, R. G. & Lecolinet, E., 1996, A survey of methods and strategies in character segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**(7), 690-706.
- Cheriet, M., Kharma, N., Liu, C. L. & Suen, C. Y., 2007, *Character recognition systems: a guide for students and practitioners*. John Wiley & Sons, Inc., New Jersey.
- Cover, T. M. & Hart, P. E., 1967, Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**(1), 21-27.
- Dougherty, G., 2013, *Pattern recognition and classification: an introduction*. Springer, New York.
- Duda, R. O., Hart, P. E. & Stork, D. G., 2001, *Pattern classification*. 2nd Ed. Wiley Interscience, New York.
- Freeman, H., 1961, On the encoding of arbitrary geometric configurations. *IRE Transactions on Electronic Computers*, **10**(2), 260-268.
- Ghosh, D., Dube, T. & Shivaprasad, A. P., 2009, Script recognition – a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(12), 2142-2161.
- Jain, A. K., Duin, R. P. & Mao, J., 2000, Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, January, **22**(1), 4-37.

- Johnson, R. A. & Wichern, D. W., 2007, *Applied multivariate statistical analysis*. 6th Ed. Prentice Hall, New Jersey.
- Kong, T. Y. & Rosenfeld, A., 1996, *Topological algorithms for digital image processing*. Elsevier, Amsterdam.
- Kozok, U., 2008, *Transtoba2*. [online] available at: <http://transtoba.seige.net> [accessed Dec 2019].
- Kozok, U., 2015, *Surat batak: Sejarah perkembangan tulisan batak, berikut pedoman menulis aksara Batak dan cap Sisingamangaraja XII*. KPG (Kepustakaan Populer Gramedia) & EFEO, Jakarta.
- Kuhl, F. P. & Giardina, C. R., 1982, Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing*, **18**(3), 236-258.
- Mu, X., Shen, X. & Kirby, J., 2017, Support vector machine classifier based on approximate entropy metric for chatbot text-based communication. *International Journal of Artificial Intelligence*, **15**(2), 1-16.
- Panggabean, M. & Rønningen, L. A., 2009, Character recognition of the Batak Toba alphabet using signatures and simplified chain code. *International Conference on Signal and Image Processing Application (ICSIPA)*, IEEE, 215-220.
- Pasaribu, N. T. & Hasugian, M. J., 2016, Noise removal on Batak Toba handwritten script using Artificial Neural Network. *3rd International Conference on Information Technology, Computer, and Electrical Engineering*, Semarang, ICITACEE, 373-376.
- Pasaribu, N. T. & Hasugian, M. J., 2016, *SoBAT (Script of Batak Toba) Database*. [online] available at: <https://aksarabatak.maranatha.edu/> [accessed Dec 2019].
- Sarma, K. K., 2009, Neural network based feature extraction for Assamese character and numeral recognition. *International Journal of Artificial Intelligence*, **2**(9), 37-56.
- Sonka, M., Hlavac, V. & Boyle, R., 2015, *Image processing, analysis, and machine vision*. 4th Ed. Cengage Learning, USA.
- Theodoridis, S. & Koutroumbas, K., 2009, *Pattern recognition*. 4th Ed. Elsevier, USA.
- Trier, O. D., Jain, A. K. & Taxt, T., 1996, Feature extraction methods for character recognition - a survey. *Pattern Recognition*, **29**(4), 641-662, .
- Yampolskiy, R. V., 2004, *Feature extraction methods for character recognition*. Master's Thesis, Dept. Computer Science, Rochester Institute of Technology, New York.
- Zheng, W., Zhao, L. & Zou, C., 2004, Locally nearest neighbor classifiers for pattern classification. *Pattern Recognition*, **37**, 1307-1309.