Myanmar Handwritten Digit and Character Recognition Using Blocked Features and Random Forest Classifier

[#1]Myint San
University of Computer Studies (Monywa)
myintsan013@gmail.com

[#2]Phyu Phyu Khaing
Myanmar Institute of Information Technology
phyu_phyu_khaing@miit.edu.mm

[#3]Moe Thida Naing
University of Computer Studies (Pakokku)
moethidanaing2018@gmail.com

Abstract: *Automatic recognition of handwritten digits and handwritten characters has been studied in the pattern recognition field for many years. The handwritten digits and character recognition are still a significant field of study, due to its wide practical applications. There has been much work in the field of Myanmar Optical Character Recognition (OCR) in recent decades. This study prepares the handwritten digit and character recognition dataset to train the model. The images in the dataset were arranged with black and white color space of 36 x 36 pixels standardized in size. This research proposed the Myanmar handwritten digit and character recognition system by using the blocked features and Random Forest Classifier. The recognition rates of Myanmar digits and characters are increased to 96.8% and 92.6% respectively with Random Forest Classifier.*

Keywords: myanmar handwritten digits and characters, blocked features, random forest classifiers

1. INTRODUCTION

Character Recognition is the recognition of written or printed text characters from input text scanned images or printed images by using technology. For many years hand written character recognition has been widely studied in the field of pattern recognition. In the marketplace, there are already many automated handwriting recognition systems. However, Myanmar handwritten character recognition is still challenging.

In Myanmar, Myanmar language is the major language for more than 54 million of Myanmar people. Myanmar language is written from left to write and it is the rounded structure language. Basically, Myanmar language is composed with ten digits and thirty-three basic characters as shown in figure 1 and 2.

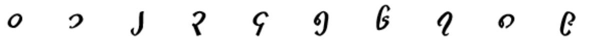


Figure 1. Ten Myanmar Digits

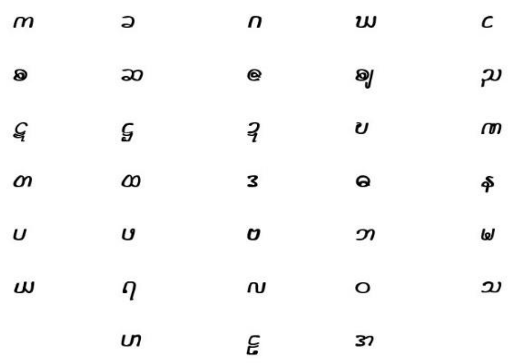


Figure 2. Thirty-three Myanmar Characters

The rest of this paper is made up of the following sections. Section 2 reviews the related work in the field of Myanmar character recognition systems. Section 3 illustrates the methodology of the system and Section 4 discusses about the experimental results in detail. Section 5 expresses some conclusions and possible future directions for science.

2. LITERATURE REVIEWS

This section discusses few important research contributions in the area of Myanmar Handwritten Recognition.

Mar et al. (2005) presented the Myanmar handwriting identification system between Exhibit and Specimen Documents. This study used the three methods to identify the Myanmar handwritten character. Three methods are Fast Fourier Transform (FFT) method for character's feature extraction, median filter for noises removal in the features, and Weighed Euclidean Distance (WED) for identification by comparing the trained features of characters. The correct identification rate is 97.5% over the forty handwritten documents [1]. Sandar (2005) introduced the off-line Myanmar handwriting and printed character recognition system. In this study, preprocessing step is used to remove the noise, recursive algorithm is used for segmentation, Hidden Markov Models classifier is applied for the recognition. The recognition rates are 92.1% and 97% for the handwriting and printed character respectively [2].

Than et al. (2006) developed the handwritten Myanmar alphabet recognition (HMAR) system. There are three main steps in HMAR system: Preprocessing, Feature Extraction and Recognition. In the preprocessing step, scanning, binarization, segmentation, noise removing, delete extra space, normalization, and thinning are processed. Zoning method are used to extract features and rule-based method is used to recognize the Myanmar alphabet. The recognition rate on the 375-training data and 125-testing data is 98.8%. This system can be extended to implement Myanmar alphabets as future work [3].

Phyu et al. (2008) introduced the on-line handwritten Myanmar recognition system for compound words based on Myanmar Intelligent Character Recognition (MICR) engine. In this system, gray scale conversion, noise filtering, binarization, extraction and normalization are processed for preprocessing. MICR engine is used to recognize the preprocessed characters and the code is arranged the recognized characters. [4]. Htwe et al. (2008) initiated the recognition model for the three items on the bank cheque written by Myanmar language such as payee’s name, legal amount and courtesy amount. Noise removal, cropping, thinning, and normalization are processed for preprocessing. Row segmentation and column segmentation are performed for the segmentation step, and MWR algorithm is used for the feature extraction step. Hidden Markov Model (HMM) is applied for the recognition. The correct final result is gotten by adding the log probabilities [5].

Thein et al. (2010) performed using MICR (Myanmar Intelligent Character Recognition) and some neural networks for ICR (Intelligent Character Recognition) and OCR (Optical Character Recognition) technology. For preprocessing, grayscale conversion, noise filtration, binarization, row column extraction, resizing and normalization are processed. And then, statistic and semantic features are extracted by using MICR from each character. In training process, neural network is used to train the features and the voting system is applied to make the final decision [6]. Htike (2013) proposed Competitive Neural Trees (CNeT) based handwritten character recognition system for Myanmar language. In preprocessing, resizing, binarization, and thinning are worked. Eighteen features are extracted by using region-based methods. For the recognition, Competitive Neural Trees (CNeT). This paper obtained the 97% of recognition rate by implementing the 33 Myanmar handwritten character for training and 330 for testing [7].

Zaw et al. (2017) presented the recognition of Myanmar Character using eight frequency chain code features based on the character as a whole and 16 blocks. For segmentation, horizontal cropping and vertical cropping methods are applied for line segmentation and character segmentation respectively. And then, 25 features are extracted. For classification and recognition, matching method is used to match the features. 92.28%, 87.34%, and 86.31% recognition rates are achieved for the isolated character, compound word and text line images discretely [8]. Khin et al. (2018) publicized Myanmar character extraction system for the license plate number. This system processed with three main steps. The first step is preprocessing by converting gray scale, binarization, and noise removal. The second step is character segmentation and extraction by constructing black and white label and region props. In the last identification step, the correct extraction rate is 90%. For the future study, this system will be implemented on Thai languages or Cambodia languages [9].

Win et al. (2019) proposed the offline Myanmar handwriting recognition system by combining deep learning. At the first step, binarizing, noise removal, segmenting and resizing is implemented for the preprocessing. Convolutional neural network is used to extract feature and recurrent neural network is used to recognize. This system showed the performance of the system with character error rate (CER) and the word error rate (WER). CER and WER are 0.24 and 0.36 for under 20 characters text and 0.27 and 0.56 for above 20 characters text [10]. Aung et al. (2019) initiated the feature extraction techniques for Myanmar characters recognition. Six important features are extracted to get the better performance. These are aspect ratio, termination points, bifurcation points, horizontal stroke point, vertical stroke point and weight direction. And then, characters are recognized by training with convolutional neural network [11].

There are many researches for the Myanmar handwritten recognition but it has still challenged to be continuously study. This study aims to learn the effective features to recognize the handwritten Myanmar Digits and Characters.

3. METHODOLOGY

This system introduced Myanmar handwritten digits and characters recognition by using the blocked features and random forest classifier. Figure 3 shows the system architecture of the system.

3.1. Preprocessing

This system implemented the six steps for the preprocessing:

1. Image Resizing,
2. Grayscale Conversion,
3. Noise Removal,
4. Image Binarization,
5. Image Inversion, And
6. Image Skeletonization.

Firstly, the input image is transformed into the fixed sized images (36 x 36). And then, RGB image is converted into the grayscale image. For the noise removal, this system used the median filter to increase the performance of the recognition process. And then, binarization is processed by adjusting the automatic threshold. If the threshold value is higher than the pixel value, the pixel value is converted into black; and otherwise, the pixel value is converted into white.

After binarization, image inversion is processed. The background of handwritten image is white and the foreground color of digits or characters shape is black. For the recognition, the background is black and the foreground is white.

Finally, skeletonization is processed to reduce the region of the binary image. for the binary image skeletonization, image skeletonization technique is the transformation technique of medial axis by iteratively deleting the boundary points of the object region of image.

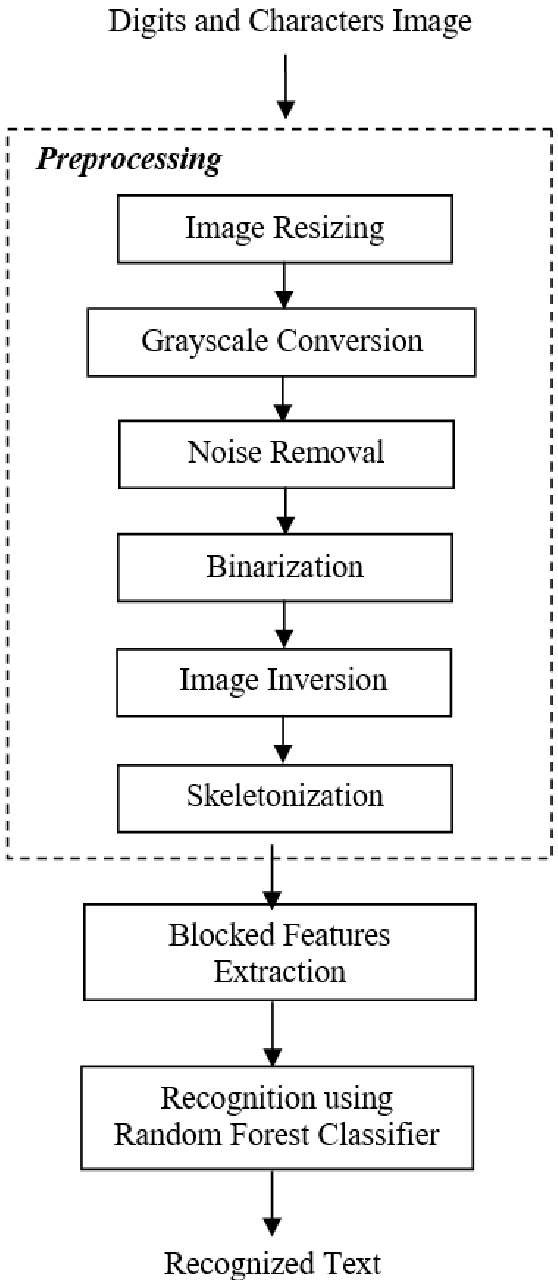


Figure 3. System Architecture

3.2. Feature Extraction

After preprocessing, features are extracted to recognize the digits or the characters of image. This system proposed the efficient feature extraction technique by blocking the image into nine regions. The size of preprocessed image is 36 x 36, and the preprocessed image is equally divided into nine regions. So, the size of one block of image is 12 x 12 as shown in figure 4.

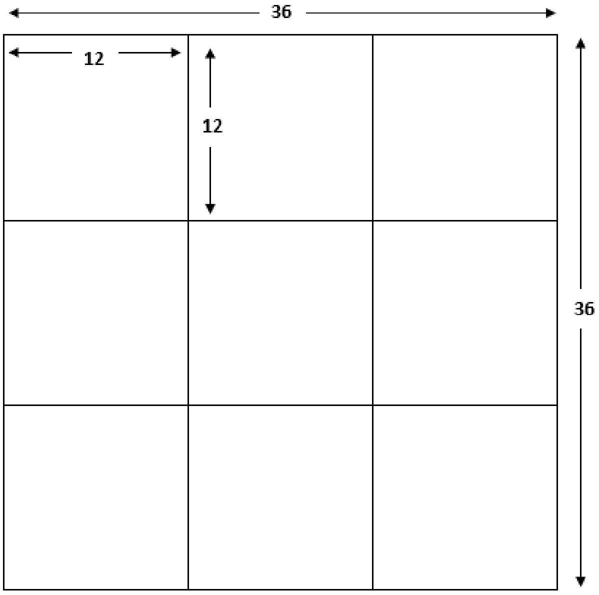


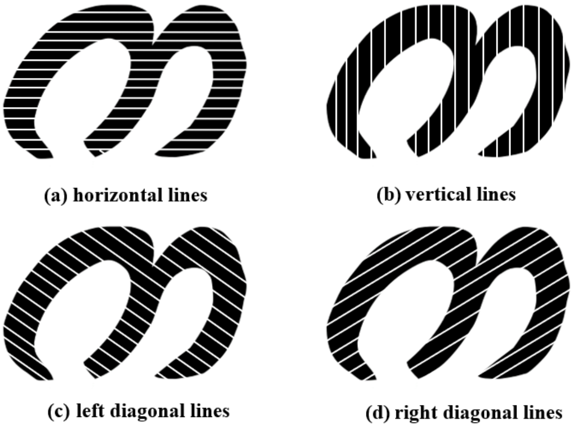
Figure 4. Image Blocked Structure

From each individual block, four important features are extracted to recognize the digits or the characters of image. Myanmar digits and characters are constructed with many horizontal lines, vertical lines, and diagonal lines. So, we decided to extract the four important features from the image. Four important features are number of lines by horizontally, vertically, right diagonal and left diagonal. Our research calculates number of horizontal lines as Figure 5(a), number of vertical lines as Figure 5(b), number of left diagonal lines as Figure 5(c), and number of right diagonal lines as Figure 5(d). Nine blocks are extracted thirty-two features to get the high performance of the recognition system.

3.2.1. Horizontal Line and Vertical Line

Horizontal line is a straight line on the plane of coordinate where the y-coordinate is the same for all points on the line. Vertical line is also a straight line on the plane of coordinate where the x-coordinate is the same for all points on the line.

Horizontal line layers may create rhythms or patterns in an image that can become a focal point of an image in itself. Horizontal lines breaking with an object or vertical lines intersecting can also create interest. So, our research focus to extract the horizontal line and vertical line as features.



(a) horizontal lines

(b) vertical lines

(c) left diagonal lines

(d) right diagonal lines

Figure 5. Image's Features Examples

3.2.2. Left Diagonal Line and Right Diagonal Line

Diagonal line is a line with slanted, or a line that connects one corner to the farthest corner. Left diagonal line is a line from the upper left corner to the lower right, and right diagonal line is a line from the lower left corner to the upper right. Regardless of the width or height of the cell, the lines will go from corner to corner.

The writing style of many people cannot be the same. Some people are writing by focusing to the right, while others are writing by focusing to the left. That's why we chose left diagonal line and right diagonal line as the features.

3.3. Recognition

In this paper, random forest classifier is used to recognize the Myanmar handwritten digits and characters from the image. Random forest classifier is the best to get the high recognition accuracy. We choose the Random Forest classifier that as the best recognition results based on thirty-two blocked features by comparing with other four classifiers, such as Multilayer Perceptron, K Nearest Neighbor, Naïve Bayes, and Support Vector Machine.

3.3.1. Random Forest Classifier

Random Forest is a ensemble learning method that processes by creating many decision trees. So that algorithm is called an ensemble tree-based learning algorithm. Every single tree in the random forest explodes out for the prediction of class and It aggregates the votes from different decision trees to decide the final class of the test object. The class that gets the most voting becomes the prediction of the model.

Random forest Classifier works with the algorithm 1 [12].

Algorithm 1. Random Forest Classifier

```
1. For $b$ = 1 to $B$:
   (a) Draw a bootstrap sample Z* of size $N$ from the training data.
   (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is reached.
       i.   Select $m$ variables at random from the $p$ variables.
       ii.  Pick the best variable/split-point among the $m$.
       iii. Split the node into two daughter nodes.
2. Output the ensemble of tree $\{T_b\}_1^B$.
3. To make the prediction at a new point x:
   Let $\hat{C}_b(x)$ be the class prediction of the $b^{th}$ random-forest tree.
   Then $\hat{C}_{rf}^B(x)$ = majority vote $\left\{\hat{C}_b(x)\right\}_1^B$
```

3.3.2. Multilayer Perceptron

Multilayer Perceptron is a feedforward artificial neural network. Multilayer feedforward neural network is shown in Figure 6. It is composed of at least three layers: an input layer, a hidden layer and an output layer. Each node, not included the input nodes, uses the nonlinear activation function. It feeds a set of inputs and generates a set of outputs by working with the backpropagation processes.

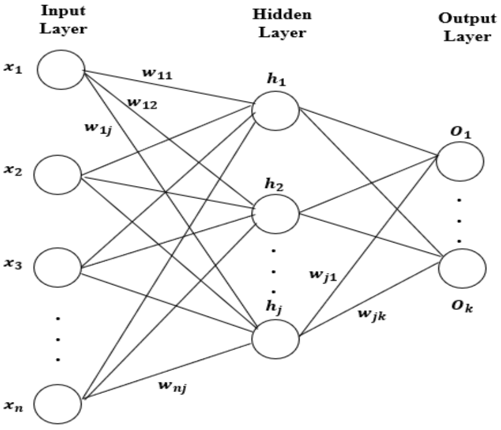


Figure 6. Multilayer Feedforward Neural Network

3.3.3. K Nearest Neighbor

K-Nearest Neighbors is a non-parametric method that used for classification. It stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). It has been used in statistical estimation and pattern recognition.

For the distance function, there are three main function: Euclidean Distance, Manhattan Distance, and ChebyShev Distance. This research is used the Euclidean Distance. Euclidean distance is working with equation 1.

$$d(x,y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (1)$$

3.3.4. Naïve Bayes

Naïve Bayes classifier is one of simple probabilistic classifiers applied the Bayes' theorem. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from prior probability of class P(c), probability of predictor given class P(x) and prior probability of predictor P(x|c) as shown in equation 2.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (2)$$

Naïve Bayes algorithm performs followed by the below steps:

Step 1. Convert the data set into the frequent table

Step 2. Create Likelihood table by finding the probabilities.

Step 3. Use Naïve Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

3.3.5. Support Vector Machine

Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression. In the SVM algorithm, each data item is plotted as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes very well.

4. EXPERIMENTAL RESULTS

This experiment is implemented the 250 images for the handwritten digits and 825 images for the handwritten characters. Although there is no standard digits or character dataset for the Myanmar handwritten image, we constructed the by using the free image editing tools from imageoneline.co website. That dataset is created twenty-five images for each of ten digits and each of thirty-three characters.

Table 1 shows the experimental results of this research. For ten Myanmar digits, Random forest classifier is 96.8%, Multilayer perceptron is 95.6%, K Nearest Neighbor is 94.4%, Naïve Bayes is 94%, and Support Vector Machine is 90%. For the thirty-three Myanmar characters, Random Forest is 92.6%, Multilayer Perceptron is 90.2%, K Nearest Neighbor is 89.8%, Naïve Bayes is 88.6%, and Support Vector Machine is 84.5%. For the accuracy, we evaluated with 10-fold cross validation on the datasets by using the Weka software.

Table 1. Experimental Results of Myanmar Handwritten Digits and Characters

Classifiers	Digits	Characters
Random Forest	96.8%	92.6%
Multilayer Perceptron	95.6%	90.2%
K Nearest Neighbor	94.4%	89.8%
Naïve Bayes	94%	88.6%
Support Vector Machine	90%	84.5%

Table 2 shows the confusion matrix for ten digits by using Random Forest Classifier. We implement 250 total images (25 images for each digits) for ten digits. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions. In the Myanmar digits (၀ to ၉), five Myanmar digits (၀ ၂ ၃ ၅ ၇) are hundred percent classified. But, four digits (၄ ၆ ၈ ၉) are ninety-six percent classified and the rest one (၁) is eighty-six percent classified.

Table 2. Confusion Matrix for Myanmar Handwritten Digits

	၀	၁	၂	၃	၄	၅	၆	၇	၈	၉
၀	25	0	0	0	0	0	0	0	0	0
၁	3	21	1	0	0	0	0	0	0	0
၂	0	0	25	0	0	0	0	0	0	0
၃	0	0	0	25	0	0	0	0	0	0
၄	0	0	0	1	24	0	0	0	0	0
၅	0	0	0	0	0	25	0	0	0	0
၆	0	0	0	0	0	0	24	0	0	1
၇	0	0	0	0	0	0	0	25	0	0
၈	1	0	0	0	0	0	0	0	24	0
၉	1	0	0	0	0	0	0	0	0	24

We implement 825 total images (25 images for each characters) for thirty-three characters. The recognition rate is 98.4% for the first ten characters, 89.5% for the second ten characters, and 95.8% for the last thirteen characters. According to the column space, we show the confusion matrix by dividing the characters. Table 3 only shows the confusion matrixes for first ten characters for Myanmar language.

Table 3. Confusion Matrix for first ten Myanmar Handwritten Characters

	က	ခ	ဂ	ဃ	င	စ	ဆ	ဇ	ဈ	ည
က	25	0	0	0	0	0	0	0	0	0
ခ	0	25	0	0	0	0	0	0	0	0
ဂ	0	0	25	0	0	0	0	0	0	0
ဃ	0	0	0	25	0	0	0	0	0	0
င	0	0	0	0	25	0	0	0	0	0
စ	0	0	0	0	0	25	0	0	0	0
ဆ	0	0	0	1	0	0	24	0	0	0
ဇ	0	0	0	0	0	0	0	25	0	0
ဈ	0	1	1	0	0	0	0	0	23	0
ည	0	0	1	0	0	0	0	0	0	24

In the handwritten Myanmar character recognition of this study, similar characters are often mistaken to recognize. Sometimes, it can also mistake when the written text is not clear. For example, if people don't conscious the head of the **“ဃ”** character when writing it, the system cannot be correctly recognized and it can wrong recognize as **“ယ”**.

Although the performances of classifiers are not very difference in the handwritten Myanmar digits, these are a little obvious in the handwritten Myanmar characters. Random Forest that get the highest performance is differ with Support Vector Machine that get the worst performance in this system. K-Nearest Neighbor and Naïve Bayes are a little differ in the performance for both two cases.

5. CONCLUSIONS

This paper proposed the Myanmar handwritten digits and characters recognition system by applying the efficient blocked features and Random Forest classifier. This research implements the ten handwritten digits and thirty-three characters of Myanmar language. The recognition results achieve 96.8% and 92.6% for ten digits and thirty-three characters for Myanmar handwritten image. The experiment only implemented the separated digits and characters and we do not consider the compound words. In the future, we will collect more handwritten images and will implement the compound words.

6. REFERENCES

[1] S.H. Mar, and N.L. Thein, “Myanmar Character Identification of Handwriting Between Exhibit and Specimen”, In 6th Asia-Pacific Symposium on Information and Telecommunication Technologies, IEEE, November 2005, pp. 95-98.

[2] K. Sandar, “A Comparison of Recognition for Off-line Myanmar Handwriting and Printed Characters”, In 6th Asia-Pacific Symposium on Information and Telecommunication Technologies, IEEE, November 2005, pp. 105-110.

[3] Y.Y. Than, D.M. Aung, A.M. Yi and K.T. Win, “Development of Handwritten Myanmar Alphabet Recognition”, International Journal of Video & Image Processing and Network Security (IJVIPNS), vol. 9, no. 10, 2006, pp. 23-29.

[4] E.E. Phyu, Z.C. Aye, E.P. Khaing, Y. Thein and M.M. Sein, “Recognition of Myanmar Handwritten Compound Words Based on MICR”, In the 29th Asian Conference on Remote Sensing (ACRS), Colombo, Sri Lanka, 2008.

[5] N.A.A. Htwe, S.S. Mon and M.M. Sein, “Recognition on User-Entered Data from Myanmar Bank Cheque”, In International Conference on Computer Application, 2008, pp. 114-118.

[6] Y. Thein, and S.S.S. Yee, High Accuracy Myanmar Handwritten Character Recognition using Hybrid approach through MICR and Neural Network”, International Journal of Computer Science Issues (IJCSI), vol. 7, no. 6, November 2010, pp. 22-27.

[7] T. Htike and Y. Thein, “Handwritten Character Recognition Using Competitive Neural Trees”, IACSIT International Journal of Engineering and Technology (IJET), vol. 5, no. 3, June 2013, pp. 352-356.

[8] K.P. Zaw and Z.M. Kyu, “Myanmar Character Recognition Using Eight Direction Chain Code Frequency Features”, International Journal of Computer and Information Engineering (IJCIE), vol. 11, no. 11, 2017, pp. 1229-1233.

[9] O. Khin, M. Phothisonothai, and S. Choomchuay, “Myanmar Character Extraction from Vehicle Images Using Aspect Ratio and Bounding Box”, In 2018 International Workshop on Advanced Image Technology (IWAIT), IEEE, pp. 1-4, 2018.

[10] H.Y. Win, and T.T. Wai, “Implementation of Myanmar Handwritten Recognition”, In International Conference on Intelligent Computing & Optimization, Springer, Cham, October 2019, pp. 320-328.

[11] Z.Z. Aung, C.M.M. Maung, and Y. Htun, “Feature Extraction in OCR for Myanmar Old Printed Documents”, International Journal of Innovative Science, Engineering & Technology (IJISET), vol. 6, no. 4, April, 2019, pp. 165-170.

[12] A. Liaw, and M. Wiener, “Classification and regression by randomForest,” *R news*, vol. 2, no. 3, 2020, pp. 18-22.