ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse83912020.pdf
https://doi.org/10.30534/ijatcse/2020/83912020

Recognition of Baybayin Symbols (Ancient Pre-Colonial Philippine Writing System) using Image Processing

Mark Jovic A. Daday[1], Arnel C. Fajardo[2], Ruji P. Medina[3]
[1] Technological Institute of the Philippines, Quezon City, Philippines, jovicmark28@gmail.com
[2] School of Computer Science Manuel L. Quezon University, Diliman, Quezon City, Philippines, acfajardo2011@gmail.com
[3] Technological Institute of the Philippines, Quezon City, Philippines, ruji.medina@tip.edu.ph

ABSTRACT

The goal of this paper is to accomplish an Optical Character Recognition (OCR) that gives an extremely contribution to the advancement of technology in terms of image recognition in Machine Learning. The researcher introduces the Feed-Forward Neural Network with Dropout Method (FFNNDM) and Convolutional Neural Network with Dropout Method (CNNDM) for the recognition of the Baybayin symbols. The phases of preprocessing of data also describe in this paper to be feed into the image recognition algorithms. The arrangement of FFNNDM is composed of one (1) dense input layer and then having a four dense (4) hidden layer and one (1) dense output layer, and the CNN structure is composed of three (3) convolutional layer, two (2) dense hidden layer and one (1) output layer. The result shows that FFNNDM is more accurate and gains an accuracy of 92.4%, loss of 0.25% and error rate of 7.55%, while the CNNDM had only an accuracy of 91.69%, loss of 0.31% and error rate of 8.31%. It also presents the confusion matrix of each algorithm to exhibit the true correct predictions of each Baybayin symbols.

Key words : OCR, Image Preprocessing, CNN, FFNN, Dropout Method

1. INTRODUCTION

Image recognition and classification is one of the interesting research areas in artificial intelligence since we place our intellectual abilities in a computer system within an algorithm in [1]. An OCR is one important field in Artificial Intelligence (AI) that gives an extremely contribution to the advancement of technology in terms of image recognition. OCR is a method of transforming digitized or manually written text or images are keen on machine-readable data that effortlessly produced by a computer system in [2]. Discrete techniques are being applied for OCR in distinctive dialects. The core phases of OCR are preprocessing, segmentations and recognition. Distinguishing manually written text is tougher than distinguishing digitized text in [3].

Many different studies had been published for the past few years but existing tools for OCR doesn't manage adequately because of the recognition problems encountered. Due to its problem faces throughout recognition, computer is incapable to take out the features correctly when scanning them in [4].

Alternatively, unusual modern methods for a several concealed datasets of text and images were faced into more experiments, that directs to a compilation of multiple type of font and unusual ruin degree in [5]. The goal of this paper is to accomplish an OCR system for the Baybayin handwriting symbols or so called ("Alibata) an ancient and national handwriting system of the Philippines. In this paper, the basic Baybayin symbols are used for study.

In this study, a manually written type recognition technique was built on the proposed fundamental method see figure 5. This study will give an impact in the area of academics and technology with the help of image processing with OCR for determining the Baybayin symbols.

The research has use OpenCV library to make the image processing of the image data. The process includes as follows: image resizing to 28x28, applying Otsu thresholding, and normalization, then the segmentation of the data. It is composed of different Baybayin symbols and divided into testing and training data see table 1. The Convolutional Neural Network with Dropout Method (CNNDM) and Feed-Forward Neural Network with Dropout Method (FFNNDM), are being used to evaluating the accuracy, loss, and error of recognizing the symbols to determine who of the two is more efficient when recognizing the Baybayin symbols.

2. REVIEW OF RELATED LETIRATURE

2.1 Baybayin

There are few studies about Baybayin symbols had been conducted by researchers but there is no work on automatic recognition of Baybayin. The Baybayin or "Alibata" became existing around year 1200's. Baybayin was also approved in the house of congress in the Philippines to be a national writing system of the Philippines with a House bill No. of 1022. The extinct symbols are classified as *syllabaries* in the study of writing systems. The Philippine *syllabaries* have 17 signs representing 14 consonants which composed of *ba, da, ka, ga, la, ha, ma, nga, na, sa, pa, wa, ya, ta,* and 3 standalone vowels composes of *a, i, and u* at some late point the

Mangyan's added *ra*. There are also three (3) different stroke used in Baybayin, the first is Tagbanwa see in figure 1, second Hanunóo see in figure 2 and third is Buhid see figure 3.

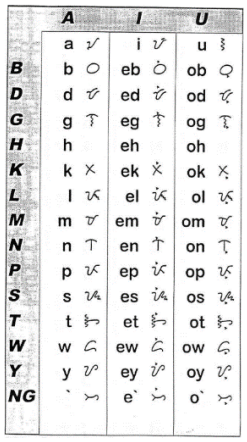

Figure 1: Quick Entry Chart of Tagbanwa

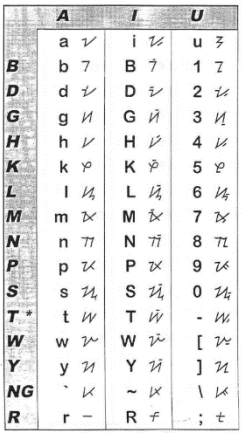

Figure 2: Mapped Entry Chart of Hanunóo

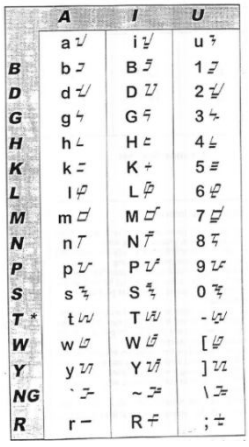

Figure 3: Mapped Entry Chart of Buhid

2.2 Optical Character Recognition

OCR was the solution in the field of AI that have matured significantly for many languages around the globe. OCR is a scheme that had the capable of identifying digital or handwritten illustrations and saved the ideal data into databases. The recognition procedures performed by means of scanning apparatuses into the illustrations and it can also perform in offline mode. OCR contains of three (3) core stages: preprocessing, segmentation, and recognition see figure 4.

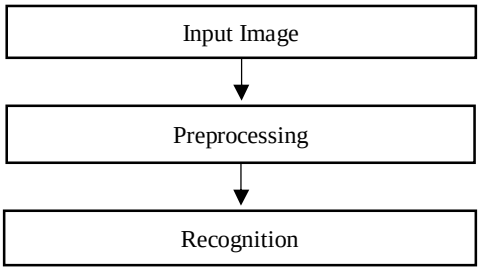

Figure 4: OCR Core Stages

Due to various applications and techniques of OCR, there's a bunch of detection and classification schemes had been suggested by several inventors and some have made an impact to the image processing field. A paper in [6] they had used the OCR and NLP that were being tested with a system into a different collection of engrave images, the implementation and accurateness of the system with regards to the disposition levels recognizing were observed. However, the accurateness of their system is varying at property of their data. Distinctively, it also varies at image resolution.

In paper [7], they've presented an image recognition of Kannada character by means of Contour Feature Extraction (CFE) technique and KNN classifier and the outcome was perceived that the accurateness of their work had a score of 84.72% after individual instants were measured to their whole dataset and they've presume that CFE technique alongside with KNN classifier could cast successfully in favor of their data then the entire properties are not counted for recognition.

Another paper [8] they've been exhibited their unique design on antique Cham glyph from the folks since 6th to 15th century, they shown how operational can ML method is useful for their research. Few quantities of data were tested and their F1-score gains 86.3% accurateness with the properties obtained from GoogleNet and KNN classifier, it also exhibiting an appropriate capable performance.

While in paper [9] they've performed connected component analysis and contour techniques in gathering the data of answer booklet and used single layer Auto Resonance Network for evaluation, their work achieved an accuracy of 91%.

3. EXPERIMENTAL

The concept of this study is composed of several parts which is shown in figure 5. This study is simulated on Anaconda3 IDE and Phyton 3.6 which is installed in an ASUS Laptop Intel Core-i5 with 3.4GHz processor and having a 4GB of RAM.

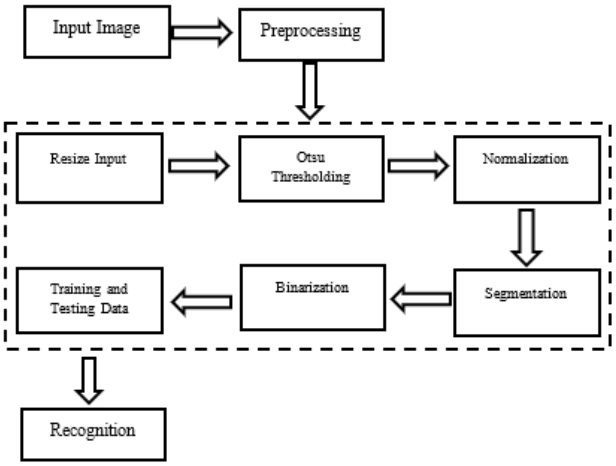

Figure 5: Steps in recognizing Baybayin Symbols

3.1 Input Data

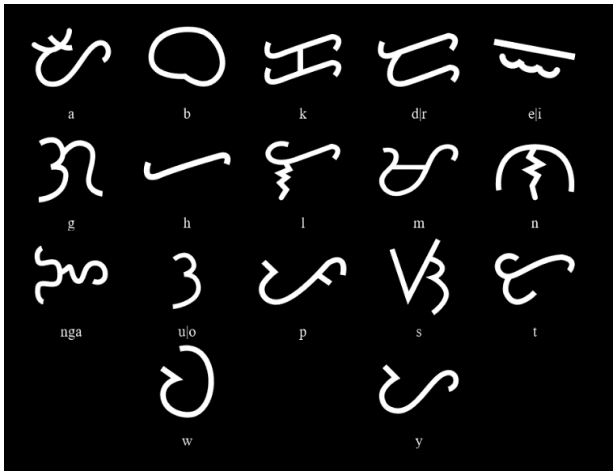

Figure 6: Baybayin character symbols

The data to be used is an existing dataset for Baybayin can be downloaded at https://www.github.com/jmbantay/Baybayin-Handwritten-Character-Dataset composing of image data sample of Baybayin see figure 6., with the combination of consonant and vowels that having of 36,000 image data see table 1.

Table 1: Baybayin Data Samples

Labels	Baybayin	Total No. of image data
0	a	1,300
1	b	2,085
2	d	2,400
3	E	1,290
4	g	2,475
5	h	2,390
6	k	2,560
7	l	2,280
8	m	2,185
9	n	2,365
10	N	2,370
11	U	1,200
12	p	2,200
13	s	2,100
14	T	2,300
15	W	2,300
16	Y	2,200
	Total	36,000

3.2 Preprocessing stage

The reason why need preprocessing of input data, is to make a neat rendering of the input data and retaining important properties that will distinguish the form of its characteristic. During the Preprocessing stage the input image are resized to 28x28 and applied by Otsu thresholding to have a clear detail of the input images, then after thresholding the normalization of data were implemented for segmentation so that to create a binarization to have a training data and testing data to be feed into recognition algorithms.

3.3 Recognition

After the preprocessing stage of data, the CNN and FFNN with Dropout Method, will be used in for recognition of Baybayin symbols, it will evaluate the accuracy, loss, and error rate of the model, also to find out what model will be suitable for recognition of Baybayin.

4. RESULT AND DISCUSSIONS

In this paper, the researcher has used the CNNDM with Dropout Method and FFNNDM. The arrangement of FFNNDM is composed of one (1) dense input layer and then having a four dense (4) hidden layer and one (1) dense output layer, and the CNN structure is composed of three (3) convolutional layer, two (2) dense hidden layer and one (1) output layer. The CNNDM with had an accuracy result of 91.69% and loss of 0.31% with an error rate of 8.31% see fig. 7 and fig. 8., however the FFNNDM and having a promising result of an accuracy rate of 92.4% and loss of 0.25% with an error rate of 7.55% see figure 9 and figure 10.

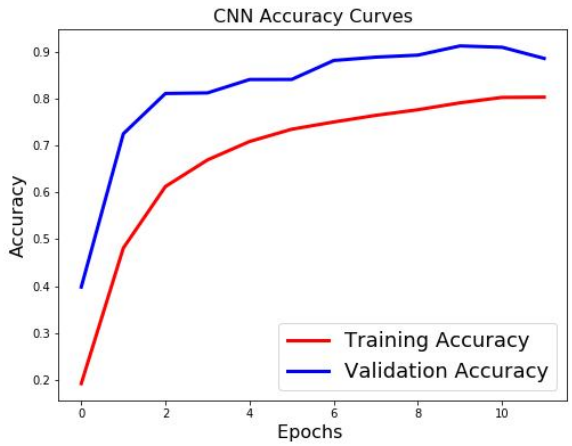

Figure 7: CNN accuracy result.

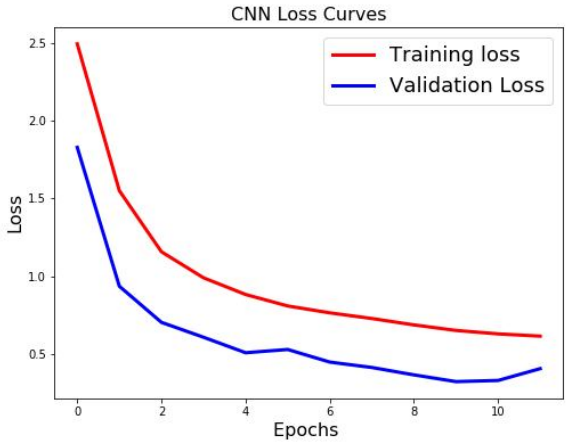

Figure 8: CNN loss result.

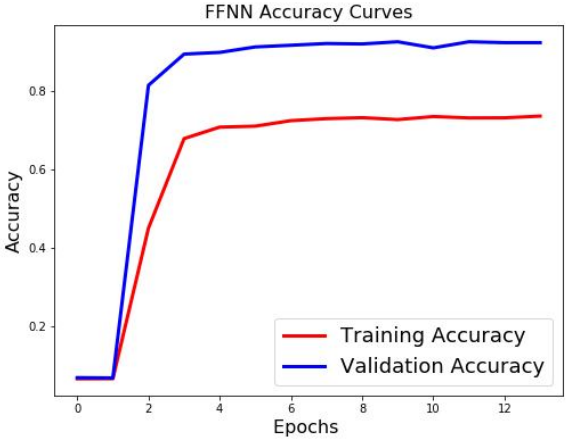

Figure 9: FFNN accuracy result.

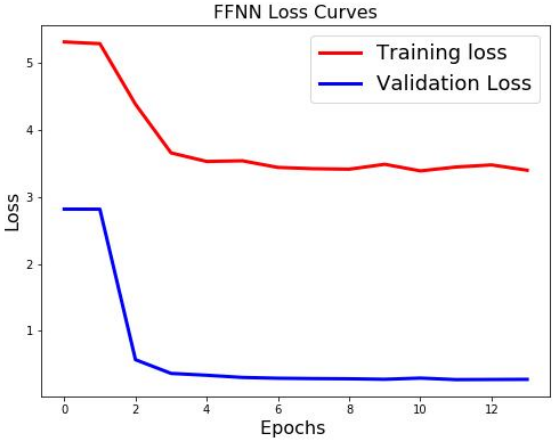

Figure 10: FFNN loss result.

It was also presented in this paper the confusion matrix for evaluation of each algorithms and show the numbers of true identified Baybayin symbols see figure 11 and figure 12.

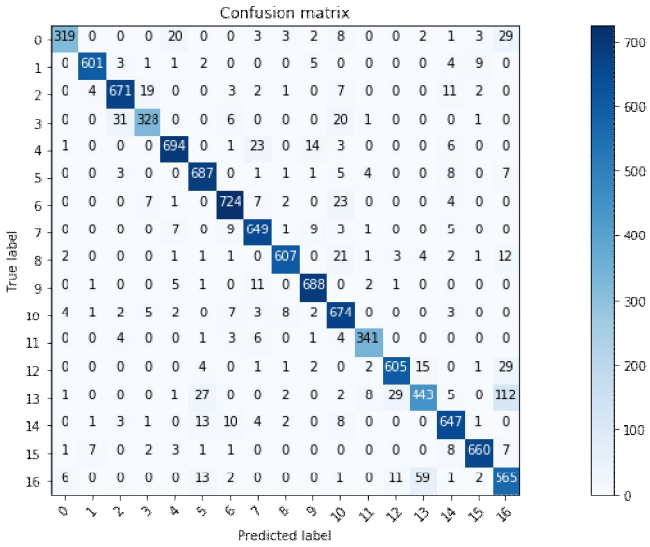

Figure 11: CNN confusion matrix of true predicted Baybayin symbols.

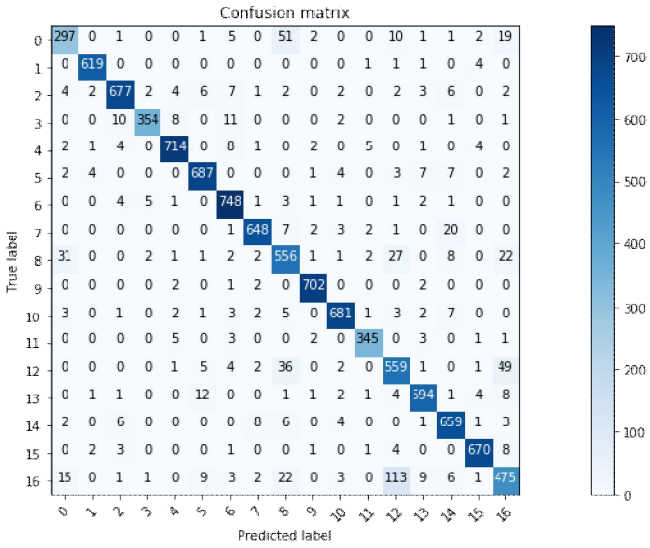

Figure 12: FFNN confusion matrix of true predicted Baybayin symbols.

5. SUMMARY

In this paper the researcher used the Baybayin Symbols an Ancient Pre-colonial Writing System in the Philippines, it includes also in this paper the preprocessing steps and two (2) image recognition algorithms that have been used and it was compared to show what algorithm will be suitable for recognizing the Baybayin sysmbols.

6. CONCLUSION

This study shows that FFNN with Dropout Method is more suitable for recognition of Baybayin Symbols than the CNN with Dropout Method in terms of accuracy, loss, and error rate. It also shows the confusion matrix of each algorithm and the FFNN is having more correct recognized Baybayin Symbols than CNN.

7. RECOMMENDATIONS

The researcher would further apply the said algorithm to a Baybayin words, sentences or even phrases for recognition with the used of the same two algorithms. The research also wants to enhanced the preprocessing stage to have a better result.

REFERENCES

1. A. B. S. I. S. W. A. Ghulam Farooque. ***Coin Recognition with Reduced Feature Set SIFT Algorithm Using Neural Network***, *y 2016 International Conference on Frontiers of Information Technology*, pp. 93-98, 2016, doi:10.1109/FIT.2016.23
2. B. M. A. E. M. F. R. A. Youssef Ouadid. ***Handwritten Tifinagh Character Recognition through a Structural Approach***, *14th International Conference on Computer Graphics, Imaging and Visualization*, pp. 50-55, 2017, doi:10.1109/CGiV.2017.15
3. A. J. C. S. Pranav P Nair. ***Malayalam Handwritten Character Recognition Using Convolutional Neural Network***, *International Conference on Inventive Communication and Computational Technologies (ICICCT 2017)*, pp. 278-281, 2017. doi:10.1109/ICICCT.2017.7975203
4. P. B. S. J. T. K. Seema Yadav. ***Word Matching and Retrieval from Images***, *International Conference on Electronics, Communication and Aerospace Technology (ICECA 2017)*, pp. 318-323, 2017, doi:10.1109/ICECA.2017.8203695
5. B. H. B. a. T. A. Habtegebrial. ***Amharic Character Image Recognition***, *2018 18th IEEE International Conference on Communication Technology*, pp. 1179-1182, 2018, doi:10.1109/ICCT.2018.8599888
6. T. V. V. D. V. N. B. Manigandan. ***Tamil Character Recognition from Ancient Epigraphical Inscription using OCR and NLP***, *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)*, pp. 1008-1011, 2017, doi:10.119/ICECDS.2017.8389589
7. S. A. S. K. P. C. H. D. M. H. R. Sneha U B. ***Image to Speech Converter – A Case Study on Handwritten Kannada Characters***, *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 881-887, 2018, doi:10.1109/ICACCI.2018.8554664
8. A.-V. S. T.-L. L. T.-H. T. a. H. V. Minh-Thang Nguyen. ***Preliminary Results on Ancient Cham Glyph Recognition from Cham Inscription images***, *2019 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pp. 1-6, 2019, doi:10.119/MAPR.2019.8743540
9. S. M. R. S. S. S. R. R. T. a. V. M. A. U Pavithra. ***OCR System for Automating Answer Script Marks using Auto Resonance Network***, *International Conference on Communication and Signal Processing, April 4-6, 2019, India*, pp. 857-861, 2019, doi:10.1109/ICCSP.2019.8698095