DEVELOPMENT OF AN IMPROVED DATABASE FOR YORUBA HANDWRITTEN CHARACTER

OLUWASHINA O. OYENIRAN[1], JOSHUA O. OYENIYI[2*], LAWRENCE O. OMOTOSHO[2], IBRAHIM K. OGUNDOYIN[2]

[1]*Department of Computer Science, Ajayi Crowther University, Oyo, Oyo State, Nigeria*

[2]*Department of Information and Communication Technology, Osun State University, Osogbo, Osun State, Nigeria*

Abstract: For improved human comprehension and autonomous machine perception, optical character recognition has been saddled with the task of translating printed or hand-written materials into digital text files. Many works have been proposed and implemented in the computerization of different human languages in the global community, but microscopic attempts have also been made to place Yoruba Handwritten Character on the board of Optical Character Recognition. This study developed a novel available dataset for research on offline Yoruba handwritten character recognition so as to fill the gaps in the existing knowledge. The developed database contains a total of 12,600 characters being made up of 70 classes from a total number of 200 writers, in which 80 % (10,500) is regarded as the training and validation dataset while the remaining 20 % (2,100) is regarded as testing dataset. The dataset is available on https://github.com/oluwashina90/Yoruba-handwritten-character-database. Hence, it is the complete and largest dataset available for Yoruba Handwritten character research.

Keywords: database, yoruba, handwritten character

1. INTRODUCTION

Handwriting character recognition has been the concern of researchers in machine learning over the past decade. Computers will now be equipped to interpret human experiences in a distinctive way by having the ability to display machine handwriting. While digitization of services continues to develop worldwide, handwriting continues to be used a lot these days for many required functions, such as writing bank cheque, among others [1].

Handwritten recognition is needed because both humans and computer should have the information readable and alternative inputs cannot be predefined. In Handwritten Recognition, two major problem domains have been established [2], which are online character recognition and offline character recognition platforms. Online Handwriting Recognition allows the captured text to be automatically translated as it is written on a special digitizer or Personal Digital Assistant (PDA), where a sensor picks up the pen-tip movements and pen-up/pen-down switching [3].

The captured handwritten text is translated into letter codes used for text processing applications [2]. Offline Handwriting Recognition requires the processing of a static representation of an article that processes an image of

* Corresponding author, email: thejoshoyeniyi@gmail.com

the article from a scanner or camera [4]. The problem of Offline Handwriting Recognition is the key persistent complexity in Optical Character Recognition (OCR) because of the variety of styles in handwriting and un-benchmarked existence of handwriting, and it typically involves language-specific techniques [5].

The major approaches used in natural language processing are syntactic and semantic analysis. Syntactic analysis, also known as parsing, identifies a text's syntactic structure and the dependence links between words, which are shown on a parse tree diagram. Parsing, Word segmentation, Sentence breaking, and Morphological segmentation are some of the syntactic approaches available. The goal of semantic analysis is to figure out what words signify. Word sense disambiguation, named entity recognition, and natural language production are some of the semantics techniques available [6].

There are several methods for handwriting recognition such as the incremental method [7], part-based methods [8], slope and slant correction [9], and the semi incremental method [10]. Other methods include line and word segmentation [11] and the ensemble method [12].

There's a need to understand the structure of such languages in order to build any language character recognition system. Several works have been suggested and implemented in the computerization of different human languages in the global community, but microscopic attempts have also been made to place Yoruba Handwritten Character on the Optical Character Recognition map. In this regard, this study seeks to develop a robust database for Yoruba Handwritten Character. The Yoruba tribe has a total population of approximately 35 million, with the majority coming from Nigeria (approximately 21 % of the population of Nigeria), 1.2 million in Benin, 0.4 million in Ghana, 0.1 million in Togo, 0.1 million in Ivory Coast, 0.2 million in Europe and 0.2 million in North America [13].

Yoruba is a tonal language consisting of seven nasal vowel-exclusive vowel sounds and eighteen consonant sounds, consisting of 25 alphabets and 3 tonal signs to differentiate between words with the same spelling but different pronunciation and meaning (three-level tones: strong, low and mid (the default tone). Each Yoruba syllable must have a minimum of one tone. It is possible to divide the Yoruba orthography into phonemes and tonemes. Phonemes can be further grouped into four main classes: consonants, oral vowels, syllabic nasals, and nasalized vowels, which are an approximation of physical expression. Yoruba uses tones in addition to the phonemes. Tonemes are the representation in tone languages of contrasting tone patterns, of which Yoruba is one. Tones are marked on vowels and syllabic nasals in Yoruba orthography using acute high tone accents (the exception is on syllabic nasals marked with a macron [14]. There are five nasalized vowels and two pure syllabic nasal vowels in the orthography [15]. This study developed a novel available dataset for research on offline Yoruba handwritten character recognition so as to fill the gaps in the existing knowledge. Offline handwriting recognition can be used to enter data and read commercial documents including checks, passports, invoices, bank statements, and receipts. Other uses include, for example, plate number recognition and collecting information from business cards into a contact list

2. REVIEW OF LITERATURE

A major problem affecting the validation of the Yoruba handwriting recognition method has been the inaccessibility of the handwritten Yoruba database. In [2] an offline handwritten Yoruba database corpus for validation of the Yoruba handwritten word recognition system (YHWR) was presented, where 50 words of medical pathology were collected from the medical pathology dictionary. The medical pathology terms were translated to their Yoruba equivalence, and 200 indigenous literate writers with suitable diacritic signs handwritten the translated words. Using 300 dpi, offline handwritten information was scanned. The corpus of the database was created; the scanned images were converted to different image formats, resolutions and image sizes to test the effect of different resolutions, formats and image sizes on the Yoruba handwritten recognition method. The digitized images were used to produce a Yoruba handwriting database that could be used to verify the method of handwritten recognition. The generated database is considered to be raw data that needs some form of preprocessing before the YHWR framework can be validated.

A dataset used for deep learning studies was also developed [16], but the downside of their work is that the data sets used were focused on Yoruba vowels and are not freely available for study. While sub-databases containing characters have also been created, the dataset used is also limited to only five characters. A dataset collected from indigenous writers was also produced [17], in which handwritten Yoruba words were retrieved. By converting

RGB to greyscale, binarization, eliminating noise, normalizing, and performing skew correction, these words were scanned and then processed. In order to quantify entropy measurements in order to enhance Yoruba recognition systems, this dataset was then used to perform experiments. The dataset used, however, was restricted because it does not cover the language's basic characters and is limited to those handwritten words.

The vectors of the feature were used to train the SVM and the words were fed for recognition into the SVM classifier. Based on recognition precision, the device was evaluated. When twenty (20) handwritten words are tested, the recognition rate varies between 66.7 %, 83.3 %, 85.7 %, 87.5 % and 100 %. In [1] a novel publicly available dataset for research on offline Yoruba handwritten character recognition was described. It contains a total of 6954 characters from a total of 183 authors, consisting of several categories, making it the largest dataset available for Yoruba handwriting research as of 2018.

3. METHODOLOGY

The first step in order to create the handwritten database is the use of a form in order to get the handwritings from several writers (see Figure 1), the data acquisition form is shown in Figure 2. This form contains all Yoruba Language Characters in both upper and lower cases and administered to two hundred (200) Yoruba Language writers making a total of 14,000 datasets. This was shared across both males and females as well as different age groups in different locations. The acquired analog data was subjected to HP 1050A Scanner, so as to convert the analog data to digital at 300 dpi. The scanned document was tentatively stored into storage facility called Yoruba Handwritten Character Gallery in .pdf extension.

YORÙBÁ LANGUAGE UPPER CASE CHARACTERS									
A	B	D	E	Ẹ	F	G	GB	I	H
J	K	L	M	N	O	Ọ	P	R	S
Ṣ	T	U	W	Y					
YORÙBÁ LANGUAGE LOWER CASE CHARACTERS									
a	b	d	e	ẹ	f	g	gb	i	h
j	k	l	m	n	o	ọ	p	r	s
ṣ	t	u	w	y					
YORÙBÁ LANGUAGE TONEME UPPER CASE CHARACTERS									
À	Á	È	É	Ẹ̀	Ẹ́	Ì	Í	Ó	Ò
Ọ̀	Ọ́	Ù	Ú						
YORÙBÁ LANGUAGE TONEME LOWER CASE CHARACTERS									
à	á	è	é	ẹ̀	ẹ́	ì	í	ó	ò
ọ̀	ọ́	ù	ú						
YORÙBÁ LANGUAGE CHARACTERS WITHOUT ASCII REPRESENTATION									
Ẹ̀	Ẹ́	Ọ̀	Ọ́	ẹ̀	ẹ́	ọ̀	ọ́		

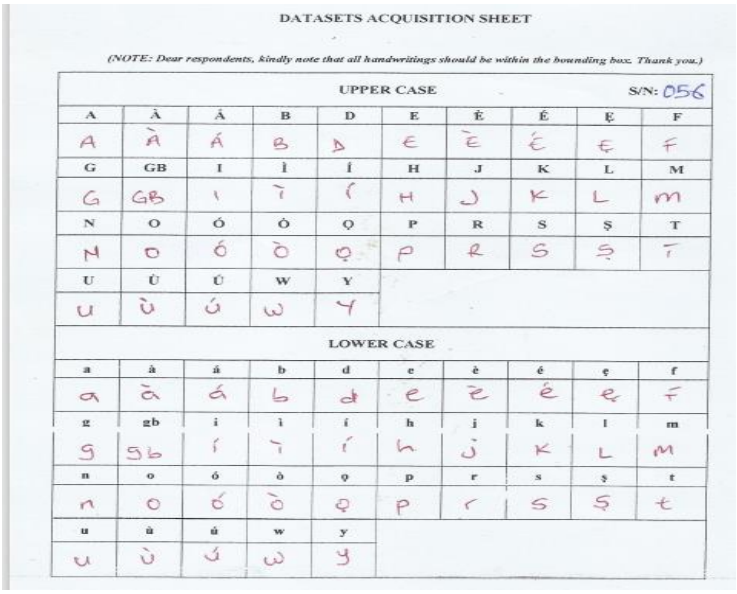
Fig. 1. Yoruba language characters [18].

In order to isolate each of the characters from the form, each filled form was subjected to the process of cropping. This process was done manually through the use of Snipping Tool software for each character. Snipping Tool is a Windows 10-based Microsoft Windows screenshot utility that can still take snapshots of an open window, rectangular areas, a free-form area, or the entire screen. You can then use a mouse or tablet to annotate the nips, store them as an image file (PNG, GIF, or JPEG file) or as an MHTML file, or e-mail them. The Snipping Tool enables simple snapshot image editing with various colored pens, an eraser, and a highlighter. The picture was saved in the .jpg file extension for the purpose of this analysis.

4. RESULT AND DISCUSSION

This study engaged two hundred writers of different age groups, gender and from different sample areas. Thus, the findings of this study with respect to Figure 3 indicate the gender distribution of the writers, where 61 % of the

writers are male while the remaining 39 % are female. This implies that majority of the writers for acquisition of dataset for this study are male.


DATASETS ACQUISITION SHEET

(NOTE: Dear respondents, Kindly note that all handwritings should be within the bounding box. Thank you.)

UPPER CASE S/N: 056

A	À	Á	B	D	E	È	É	Ẹ	F
G	GB	I	Ì	Í	H	J	K	L	M
N	O	Ó	Ò	Ọ	P	R	S	Ṣ	T
U	Ù	Ú	W	Y					

LOWER CASE

a	à	á	b	d	e	è	é	ẹ	f
g	gb	i	ì	í	h	j	k	l	m
n	o	ó	ò	ọ	p	r	s	ṣ	t
u	ù	ú	w	y					

Fig. 2. Data acquisition form.

Also, as shown in Figure 3, the age distribution of writers is as follows: 49 % are between the ages of sixteen and thirty, 35 % are between the ages of three and fifteen, and the remaining 16 % are 31 years and older. This means that the majority of the writers are aged sixteen to thirty. Figure 4 depicts the sampling area for this study. 57.5 % of the writers are Federal College of Education [Special] students, 30 % are Koladaisi University students, 7.5 % are Ajayi Crowther University students, and the remaining 5 % are Divine Intervention Nursery and Primary School students.

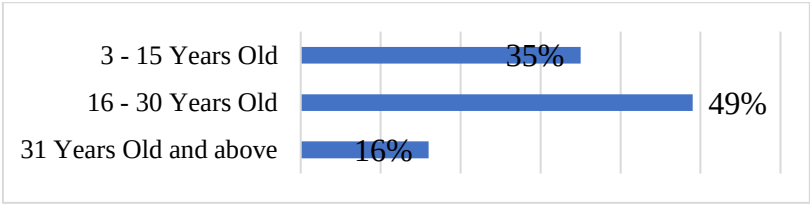


Fig. 3. Age Distribution of the writers.

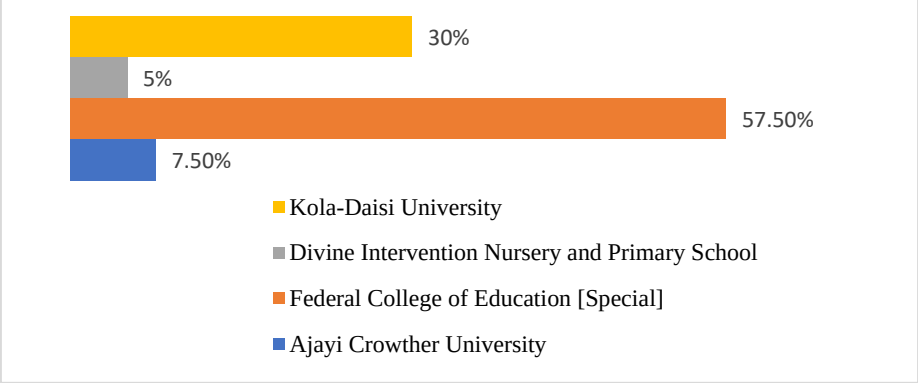


Fig. 4. Sample area.

After going through all of the data cleaning procedures, 75 % of the acquired data from the writers is valid, while the remaining 25 % is not. As a result, only the data that are 75 % valid are used in this study. The 75 % valid data was submitted to a sorting technique, in which each isolated character was sorted into its own group, and so sorted alphabetically. The total number of samples gathered and valid in this study is 12,600. Where 80 % (10,500) of

the valid datasets were put aside for training and validation, and the remaining 20 % (2,100) were isolated for further verifications. Since there are 70 classes in the dataset, it implies that for the training and evaluation there are 150 data in each class where there are 30 data in each class for the verification dataset as shown in Table 1.

Table 1. Dataset analysis.

Dataset	No. of data sample	No. of data in a class	Percentage
Training	10,500	150	80%
Testing	2,100	30	20%
Total No of dataset	12,600		

5. CONCLUSION

This study developed a locally generated datasets comprises of all the Yoruba alphabets, it acquired 12,600 samples from two hundred writers (10,500 was tagged training and validation dataset while 2,100 was tagged testing dataset). The dataset also comprises of all the tonal alphabets, thus, make it suitable for the development of any recognition or predictive system with respect to Yoruba alphabets. The dataset is made available at https://github.com/oluwashina90/Yoruba-handwritten-character-database for public use.

The developed dataset by this study is far and more advanced than the one developed in [19] whose developed system was experimented with 600 handwritten images for Yoruba alphabets, 480 of which was used for training and 120 was used for testing. Also, it is worthy of mention that this developed dataset is also more voluminous and with rich qualities as compared to the one developed in [1] which contained a total of 6954 characters being made up of several categories from a total number of 183 writers. The enhanced database developed can be used for research on sentiment analysis, text extraction, language translation, predictive text, and text analysis. It can be used to create social media monitoring apps, OCR scanners, extraction apps, digital mobile based OCR apps, chat bots, and virtual assistant apps, among others. The database can also be integrated into existing applications to improve effectiveness and efficiency by processing more data faster and with fewer resources.

REFERENCES

[1] Ojumaha, S., Sanjay, M., Adewole, A database for handwritten yoruba characters, Data science and analytics communications in computer and information, Science Springer, 2018.
[2] Ajao, J.F., Olawuyi, D.O., Odejobi, O.O., Yoruba handwritten character recognition using freeman chain code and k-nearest neighbor classifier, Jurnal Teknologi dan Sistem Komputer, vol. 6, no. 2, 2018, p. 129-134.
[3] Femwa, O.D., Development of a writer- independent online, handwritten character recognition system using modified hybrid neural network model, PhD. Thesis, Ladoke Akintola University of Technology, Ogbomoso, 2012.
[4] Ajao, J.F., Olabiyisi, S.O., Omidiora, E.O., Okediran, O.O., Hidden markov model approach for offline Yoruba handwritten word recognition, British Journal of Mathematics & Computer Science, vol. 18, no. 6, 2016, p. 1-20.
[5] Manoj, S., Narendra, S.A., Survey on handwritten character recognition (HCR) techniques for English Alphabets, Advances in Vision Computing, An International Journal (AVC), vol. 3, no. 1, 2016, p. 1 -3.
[6] Charles, C.T., Sung-Hyuk, C., English language handwriting recognition interfaces in text entry systems, 2007, http://csis.pace.edu/ctappert/papers/2007BookChap.pdf.
[7] Phan, K.M., Cuong, T.N., Anh, L.D, Masaki N., An incremental recognition method for online handwritten mathematical expression, 3rd IAPR Asian Conference on Pattern Recognition, 2015, p. 225-228.
[8] Song, W., Seiichi, U., Marcus, L., Comparative study of part-based handwritten character recognition methods, 11th International Conference on Document Analysis and Recognition, ICDAR, 2011, p. 814-818.
[9] Gupta, J., Chanda, B., Novel methods for slope and slant correction of off-line handwritten text word, Proceedings - 2012, 3rd International Conference on Emerging Applications of Information Technology, EAIT 2012, p. 295-298.
[10] Phan, K., Minh Anh, L.D, Masaki N., Semi-incremental recognition of online handwritten mathematical expressions, 15th International Conference on Frontiers in Handwriting Recognition, 2016.
[11] Banumathi, L.K., Jagadeesh, C.P., Line and word segmentation of Kannada handwritten text documents using projection profile technique, International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2016.

[12] Gunter, S., Horst, B., Creation of classifier ensembles for handwritten word recognition using feature selection algorithms, Frontiers in Handwriting Recognition, The 8th International Conference on Frontiers in Handwriting Recognition, 2002, p. 183–188.
[13] Abdulkareem, Z.O., Edet, E.E., Yor CALL: Improving and sustaining YORUBA language through a practical iterative learning Approach, OcRI, 2016, p. 1-5.
[14] Asahiah, F.O., Odẹ́jobí, O.A., Adagunodo, E.R., Restoring tone-marks in standard Yoruba electronic text: improved model, Computer Science, vol. 18, n. 3, 2017, p. 301–315.
[15] Bamgbose, A., A grammar of Yoruba, West African language monograph series, Cambridge University Press, 2014.
[16] Oyedotun, O.K., Olaniyi, E.O., Khashman, A., Deep learning in character recognition considering pattern invariance constraints, International Journal of Intelligent Systems and Applications, vol. 7, no. 7, 2015, p. 1.
[17] Ajao, J.F., Olabiyisi, S.O., Omidiora, E.O., Yoruba handwriting word recognition quality evaluation of preprocessing attributes using information theory approach, International Journal of Applied Information Systems, vol. 9, no. 1, 2015, p. 18–23.
[18] Field Survey, 2020, https://forumea.org/resources/data-collection/2020-state-of-the-field-survey/ (91.01.2021).
[19] Oladele, M.O., Adepoju, T.M., Olatoke, O.A., Ojo, O.A., Offline yoruba handwritten word recognition using geometric feature extraction and support vector machine classifier, Malaysian Journal of Computing, vol. 5, no. 2, 2020, p. 504 – 514.