A Database for Handwritten Yoruba Characters

Samuel Ojumah, Sanjay Misra(✉), and Adewole Adewumi

Covenant University, Ota, Ogun State, Nigeria
{samuel.ojumah, sanjay.misra,
wole.adewumi}@covenantuniversity.edu.ng

Abstract. This paper describes a novel publicly available dataset for research on offline Yoruba handwritten character recognition. It contains a total of 6954 characters being made up of several categories from a total number of 183 writers thus making it the largest available dataset for Yoruba handwriting research. It can be used for designing and evaluating handwritten character recognition systems for the Yoruba language as well as provide valuable insights through writer identification. The dataset has been partitioned into training and test sets being shared into 70% and 30% respectively.

Keywords: Database · Character recognition · Yoruba

1 Introduction

Handwriting character recognition has been the interest of machine learning researchers for several decades (Yadav and Yadav 2015). By being able to teach computers handwritings, computers can now be able to understand human interactions in a unique way. Although digitization of services across the globe continues to grow, handwriting is still very much in use today for many important functions such as writing bank cheques. Hence, the need for further research into this area.

The development of character recognition systems hinges on the pace at which databases can be developed and maintained with the best character sets possible required to cover the various languages. Standard handwriting databases have been developed in languages such as English (Marti and Bunke 2002), Chinese (Liu et al. 2011), French, and Arabic (Djeddi et al. 2014) as systems developed have yielded outstanding results. The advent of deep learning in which a feature learning framework for neural networks was developed has led to outstanding successes in image recognition which have been applied to these handwritten datasets with near human-level accuracy (Pascanu et al. 2013).

Handwriting recognition consist of two types: offline and online recognition. In offline recognition, the source of the input is usually a scanner that captures the character to be captured while online refers to inputs from pen or touch based inputs (Graves and Schmidhuber 2009).

The Yoruba language has a worldwide audience of about 18 million speakers with core geographic regions in Nigeria (Oyedotun et al. 2015), Togo and Benin. It is also

B. Panda et al. (Eds.): REDSET 2017, CCIS 799, pp. 107–115, 2018.
https://doi.org/10.1007/978-981-10-8527-7_10

spoken in as far as Brazil, the Caribbean and Cuba. Although, there have been several attempts to develop handwritten character recognition systems, most databases that have been used have been privately held datasets with most not being able to cover the basic character sets of the language. In order to develop character recognition systems that achieve human-level accuracy, the development of standard datasets for the Yoruba language is of paramount importance.

This work introduces a novel Yoruba handwriting dataset for character recognition which integrates features such as writer identification having fields such as the gender and age of the individual. The dataset is publicly available at https://github.com/samuelcesc/yohcrdb. The rest of this paper is as follows: "Related Works" highlights the different databases that have been developed for other languages, "Methodology" describes the processes involved in creating the dataset while "Dataset Statistics" section provides a quantitative analysis of the dataset and the "Conclusion and Future Work" section concludes the paper and gives recommendations for future improvements of this work.

2 Literature Review

This section covers the related works that have been done in the field of handwriting recognition with emphasis on handwriting databases. A review of popular languages in Africa is performed based on the context of this research as well as works done for the Yoruba language.

A major script used on the continent is the Arabic script, which is used for languages such as Swahili. The Swahili language has over 100 million speakers around Africa and is the official language for Tanzania, Kenya and Uganda. It is also spoken in nations such as Rwanda, Sudan, Ethiopia, Burundi and Comoros Islands. (Mahmoud et al. 2014) introduced an open Arabic offline text database made up of 1000 handwritten forms from 1000 different writers. The novel approach to the writer gathering was the population was from different countries. The forms containing the handwritings were scanned in various resolutions: 200, 300 and 600 dpi. These writers varied by a number of criteria such as age, upbringing country, qualification and left/right-handedness. The entire database consists of 2000 text paragraphs which generated 9000 lines and were randomly grouped into different sets for training, validation and testing. Experiments were also conducted on this data set but results show comparatively low recognition rates.

(Saabni and El-Sana 2013) presented a novel approach for generating large datasets in which prototypes for each word can be automatically generated using multiple appearances of the character. This works also shows that offline databases can be generated from online strokes based using standard dilation techniques. Experiments performed showed that the system can efficiently generated large datasets with quality as good as those involving human writers.

Another major language is Amharic, which is the official language of Ethiopia and is the second most spoken language in the country with over 21 million speakers. (Assabie and Bigun 2011) introduced a database of handwritten Amharic documents containing a total of 10,932 distinct word samples were gathered from 177 writers.

The data was collected using a form where each writer was made to write characters without any constraint. A total of 307 samples were collected and scanned at a resolution of 300dpi. Another set of 152 writers were made to write a total of 120,840 isolated character samples in which 256 characters were written thrice by each individual. This dataset was then used for training by a recognition system using the hidden Markov model.

The French language has about 120 million speakers on the African continent with core countries such as Gabon, Mauritius, Cote d'Ivoire, Senegal, São Tomé and Principe, Tunisia, Guinea, Seychelles, Democratic Republic of Congo, and Equatorial Guinea. (Djeddi et al. 2014) created a database containing 62600 French words, 74700 Arabic words, 21000 digits and 1300 signatures which were retrieved from 1300 forms from 100 writers. These forms were scanned using a scanner at 300dpi as characters were extracted and placed in categories. The writers of the handwritten words were from different ages, gender and backgrounds in terms of education which makes the database a good fit for research into writer identification and writer demographic classification.

Amazigh or Berbers is another dialect with majority of speakers in countries such as Algeria, Egypt, Tunisia, Libya, Mali, and Morocco which amount to approximately 30 million speakers of the language (Bentayebi et al. 2014). (Saady et al. 2011) created a database of handwritten Amazigh characters, which included more than 25000 characters written by 60 different writers. The main aim of this work was to create a dataset to be used for Amazigh recognition research by providing training and testing sets so that models could be developed. The samples were collected through forms and scanned at a resolution of 2400dpi and further processed by extracting the characters, which were isolated texts. These isolated texts make up the database, which is available for academic uses and further research.

(Bencharef et al. 2015) created a dataset of handwritten Tifinagh characters which is the official language of the Amazigh language. This dataset was made up of 1376 image characters written by 30 writers, which were 17 males and 13 females. These pages of the 32 handwritten Tifinagh characters were then scanned and extraction was done. The extraction process was done by using horizontal histogram in order to correct the inclination of every page and then the connected components algorithm was applied so as to detect the center of each character. The next step was then to extract 31 sub-images of 30 × 30px that contained the characters and then lastly these characters were named using Latin character set. This dataset provides a platform for efficiently training character recognition systems for Tifinagh handwriting.

(Marti and Bunke 2002) created an English sentence database for offline handwriting recognition based on the Lancaster/Oslo corpus. It consists of full English sentences produced by 1,066 forms and produced by 400 different writers most of which originate from Switzerland. A total of 9,157 lines of text can be used for text recognition while 128 lines contain information not included in the corpus used. The database contains 82227 word instances out of a vocabulary of 10,841 words and is a standard database used for English handwriting research.

As regards the Yoruba language, (Oyedotun et al. 2015) created a dataset used for experiments for deep learning. The drawback of this work is that the dataset used were based on Yoruba vowels and are not publicly available for research. The dataset used is also limited with just five characters although sub-databases containing the characters were also developed.

(Ajao et al. 2015) created a dataset acquired from indigenous writers in which Yoruba handwritten words were retrieved. These words were then scanned and then processed by conversion of RGB to greyscale, binarization, removing noise, normalizing, and performing skew correction. This dataset was then used to conduct experiments in order to calculate entropy measure so as to improve Yoruba recognition systems. However, the dataset used was limited as it does not cover the basic characters of the language and limited to certain handwritten words.

From the related works above, studies show that many standard databases provide a platform for conducting state-of-the-art research in character recognition in which features such as writer-identification can be integrated. Also, there is no publicly available database for Yoruba language with which recognition systems can be trained. In order to serve the over 30 million speakers of the Yoruba language, the creation of a publicly available dataset is of utmost importance. This work therefore addresses this gap as well as integrates other features found in many standard handwriting databases.

3 Methodology

This section describes the processes taken in order to create the Yoruba handwritten dataset.

3.1 Character Set

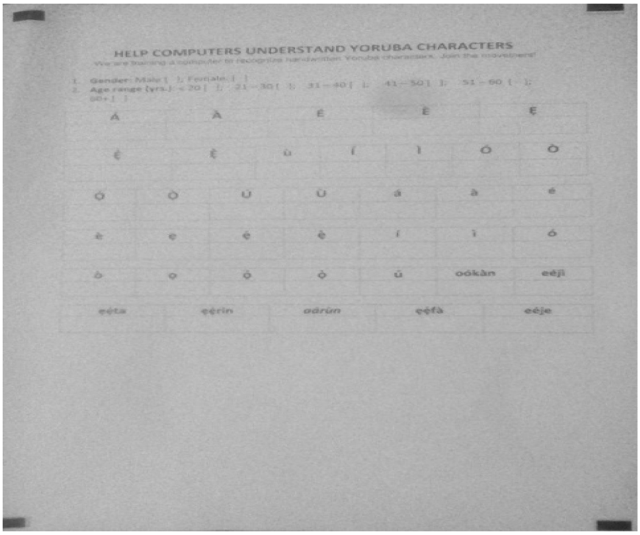
The Yoruba Character set used in this paper is made up of 31 characters and 7 words which are the words covering numbers one to seven. The stand-alone characters consist of characters with diacritic marks used in the language. Table 1 shows the character set used in the creation of the dataset.

Table 1. Yoruba Character Set

S/N	Character sets	Amount
1	Á À É È Ẹ Ẹ Ẹ ù Í Ì Ó Ò Ợ Ọ Ú Ù á à é è ẹ ẹ́ ẹ̀ í ì ó ò ọ ọ́ ọ̀ ú	31
2	oókàn eéjì ẹẹ́ta ẹẹ́rin aárùn ẹẹ́fà eéje	7

3.2 Data Collection

The first step in order to create the handwritten database was the use of a form in order to get the handwritings from several writers. To do this, literate Yoruba speakers were engaged in order to design the form to be filled. This form contained characters specific to the Yoruba language with the diacritic marks as those similar to the English language were left out. After design of the form, these forms were given to 200 writers shared across both males and females as well as different age groups. The form consisted of a total of 31 characters having fields that showed the characters and a blank



HELP COMPUTERS UNDERSTAND YORUBA CHARACTERS

We are training a computer to recognize handwritten Yoruba characters. Join the movement!

1. Gender: Male [ ]; Female [ ]
2. Age range [yrs.]: < 20 [ ]; 21 – 30 [ ]; 31 – 40 [ ]; 41 – 50 [ ]; 51 – 60 [ ]; 60+ [ ]

Á	À	É	È	Ẹ		
Ẹ́	Ẹ̀	ù	Í	Ì	Ó	Ò
Ọ́	Ọ̀	Ú	Ù	á	à	é
è	ẹ	ẹ́	ẹ̀	í	ì	ó
ò	ọ	ọ́	ọ̀	ú	oókàn	eéjì
eéta	ẹẹ́rin	aárùn	ẹẹ́fà	eéje		

Fig. 1. Blank form consisting of the Yoruba characters

space in a box beneath each character to be filled by the writer. Figure 1 shows a blank form used in data capture of the characters.

3.3 Form Processing

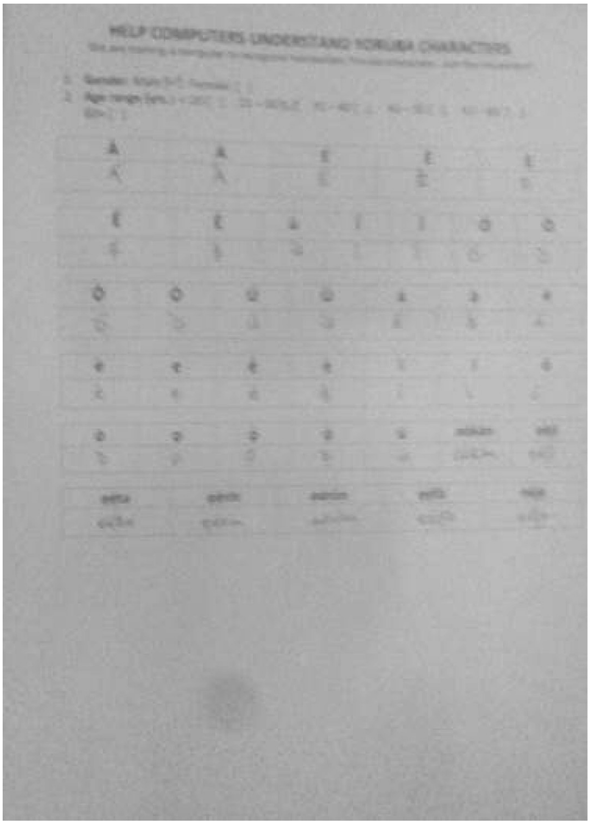
After gathering the forms, the scanning process involved using a high-quality scanner using 300dpi to capture all the forms for further processing. The varying formats such as pdf, jpeg, and bmp were made available but the bmp format was chosen for the characters. The writers captured the characters using pens with black and blue. Figure 2 shows a filled form from one of the writers containing character in each field.

3.4 Character Extraction

In order to extract the characters from the form, each filled form went through the process of cropping where each character block was cropped. This process was manually engineered using the Microsoft Paint software. Table 2. Shows a sample set of cropped characters retrieved from a form.

3.5 Pre-processing

As one of the most important aspects of setting up a database for recognition systems, pre-processing of the images was done in order to improve the quality such that recognition systems could extract features properly.



HELP COMPUTERS UNDERSTAND YORUBA CHARACTERS

1. Gender: [illegible]
2. Age range [illegible]

Fig. 2. A filled form showing handwritten characters.

Table 2. Sample of cropped characters

Ò	Ọ́	É	à	Á	Ì	Ú	Á	é	á
ẹ̀	[illegible]	Ù	OÓK	eéje	ẹẹ́t	aárù	ẹẹ́ri	ẹẹ́fà	ee

4 Dataset Statistics

From the sample set of 184 persons, a number of 183 were selected after abnormal samples were removed.

4.1 Writers Distribution

All the writers come from Covenant University, Nigeria with most being undergraduate or graduate students. Among the 183 writers, 129 (73%) are male while 54 are female. As regards the ages of the individuals, most of the writers are between the ages 21 and 30. Some writers also refused to disclose their ages. Table 3 gives a more detailed description of the age distribution of the writers. Figure 3 shows a graphic representation of the gender distribution of the writers.

Table 3. Age distribution of writers

Age	#Persons	Percentage
<20	64	34.97%
21–30	68	37.16%
31–40	2	1.09%
>40	1	0.55%
Uncategorized	48	24.59%

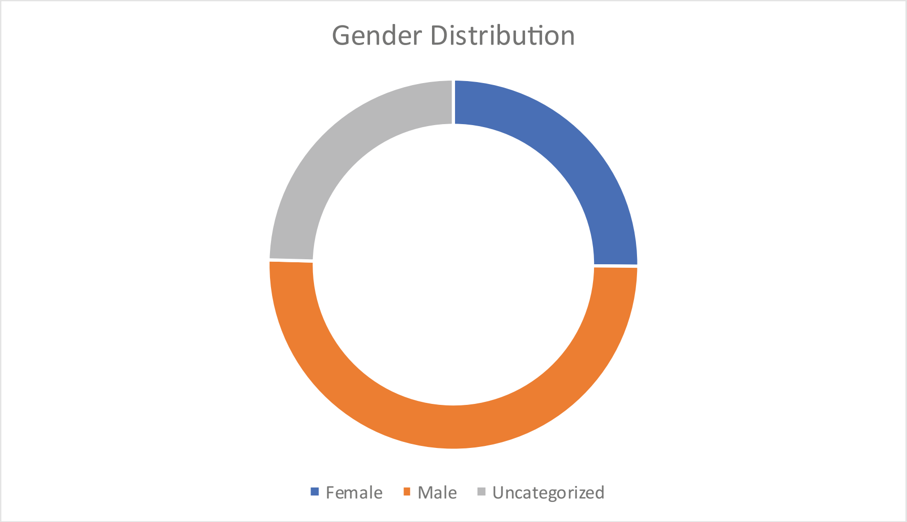


Fig. 3. Gender distribution of writers

Figure 4 shows a graphic representation of the age distribution of the writers.

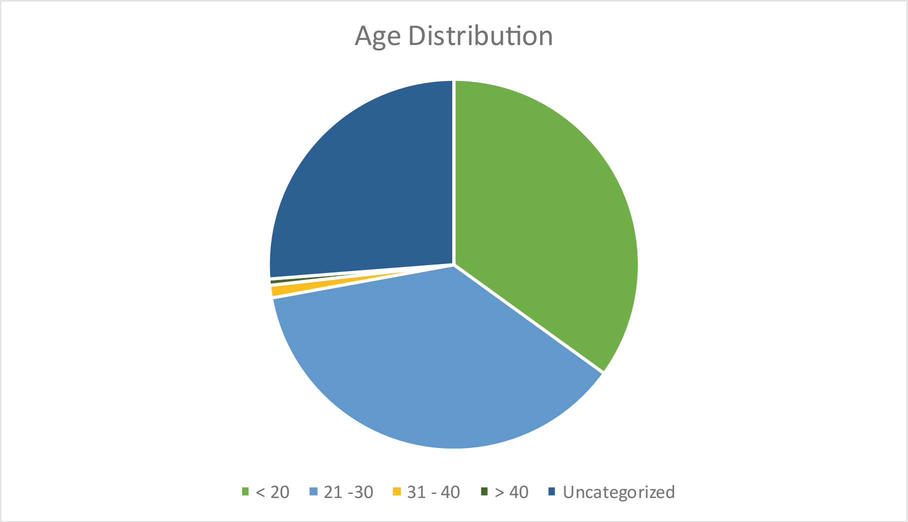


Fig. 4. Age distribution of writers

4.2 Abnormal Samples

Apart from the one sample removed from the sample set, there were no unusual samples from the 183 writers. The sample set removed contained strokes without letters of the language, which renders it, unfit for the dataset.

5 Conclusion and Future Work

A novel Yoruba offline handwritten database (YRHCR) written by several writers with support for writer identification is introduced in this work. A total of 6954 handwritten character images were gotten from 183 writers which were both male and female. Another approach to the creation of the dataset was the gathering of age ranges of the writers. These scanned images had noise removed in order to provide a platform for a clean dataset. The dataset derived from this work is the first publicly available dataset for handwritten research to be used around the world and can be used for handwriting recognition, and writer identification.

However, the database provides a platform for further work to be done as extending the database is of utmost importance in order to create a more comprehensive database that can be used for extensive research into the language in order to create tools that fit in perfectly in various use cases.

Acknowledgements. This generation of this database was done with help from Learnd Technologies, which helped from the design phase to the scanning phase. The authors thank all members of the Learnd team for the collaboration in the creation the dataset. The financial support of Covenant University Centre for Research Innovation and Discovery (CUCRID) is also acknowledged.

References

Ajao, J.F., Olabiyisi, S.O., Omidiora, E.O.: Yoruba handwriting word recognition quality evaluation of preprocessing attributes using information theory approach. Int. J. Appl. Inf. Syst. (IJAIS) **9**(1), 18–23 (2015)

Assabie, Y., Bigun, J.: Offline handwritten Amharic word recognition. Pattern Recogn. Lett. **32** (8), 1089–1099 (2011)

Bencharef, O., Chihab, Y., Mousaid, N., Oujaoura, M.: Data set for Tifinagh handwriting character recognition. Data. Brief **4**, 11–13 (2015)

Bentayebi, K., Abada, F., Ihzmad, H., Amzazi, S.: Genetic ancestry of a Moroccan population as inferred from autosomal STRs. Meta Gene **2**, 427–438 (2014)

Djeddi, C., Gattal, A., Souici-Meslati, L., Siddiqi, I., Chibani, Y., El Abed, H.: LAMIS-MSHD: a multi-script offline handwriting database. In: 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 93–97. IEEE (2014)

Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: Advances in Neural Information Processing Systems, pp. 545–552 (2009)

Liu, C.-L., Yin, F., Wang, D.-H., Wang, Q.-F.: CASIA online and offline Chinese handwriting databases. In: 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 37–41. IEEE (2011)

Mahmoud, S.A., Ahmad, I., Al-Khatib, W.G., Alshayeb, M., Parvez, M.T., Märgner, V., Fink, G. A.: KHATT: an open Arabic offline handwritten text database. Pattern Recogn. **47**(3), 1096–1112 (2014)

Marti, U.-V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. Int. J. Doc. Anal. Recogn. **5**(1), 39–46 (2002)

Oyedotun, O.K., Olaniyi, E.O., Khashman, A.: Deep learning in character recognition considering pattern invariance constraints. Int. J. Intell. Syst. Appl. **7**(7), 1 (2015)

Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. ICML **3**(28), 1310–1318 (2013)

Saabni, R.M., El-Sana, J.A.: Comprehensive synthetic Arabic database for on/off-line script recognition research. Int. J. Doc. Anal. Recogn. (IJDAR) **16**(3), 285–294 (2013)

Saady, Y.E., Rachidi, A., Yassa, M., Mammass, D.: AMHCD: a database for amazigh handwritten character recognition research. Int. J. Comput. Appl. **27**(4), 44–48 (2011)

Yadav, P., Yadav, N.: Handwriting recognition system-a review. Analysis, **114**(19), 36–40 (2015)