EVALUATION OF AN OPTICAL CHARACTER RECOGNITION MODEL FOR YORUBA TEXT

[1] Abimbola Akintola, [2] Tunji Ibiyemi, [3]Amos Bajeh

[1,3] Department of Computer Science, University of Ilorin, Nigeria
[2] Department of Electrical Engineering, University of Ilorin, Nigeria

Corresponding Author: Abimbola Akintola, akintola.ag@unilorin.edu.ng

ABSTRACT: The optical character recognition (OCR) for different languages has been developed and in use with diverse applications over the years. The development of OCR enables the digitization of paper document that would have been neglected over a period of time as well as serving as a form of backup for those documents. The system proposed is for isolated characters of Yoruba language. Yoruba language is a tonal language that carries accent on the vowel alphabets. The process used involves image gray scal, binarization, de-skew, and segmentation. Thus, the OCR enable the system read the images and convert them to text data. The proposed model was evaluated using the information retrieval metrics: Precision and Recall. Results showed a significant performance with a recall of 100% in the sample document used, and precision results that varies between 76%, 97%, and 100% in the sample document.
KEYWORDS: recognition, binarization, image digitization, accuracy, Yoruba language.

1.0 INTRODUCTION

Human Computer Interaction (HCI) is a form of communication which entails the study, planning and design of communication between people and computers. ([Kar08]). Optical Character Recognition (OCR) is a form of HCI and the technology enables the electronic or mechanical conversion of different types of documents by digitization into readable, editable and searchable document, such as scanned documents, images captured by a digital camera or PDF files. Character is used to build structures of a language and it is the fundamental building block of any language. These are the alphabets and the structures are strings, words and sentences ([CK14]). Yorùbá language is indigenous to Nigerians (where it is an official language), Togolese and Benin republicans. It is used in interaction by about 50 million people in south west Nigeria and across the world. ([A+14])
Yorùbá is a tonal language like many African languages and it is one of the Niger-Congo language families. Therefore, the meaning of a word is in the tone. Sounds in many languages are produced when alphabets are combined together. In tonal languages like Yorùbá, words of the same spelling can have different meanings. These words are called homographs. Hence, it is the tonal sign placed on them that distinguishes their meaning and not their spelling
Yorùbá alphabet can be divided into consonants and vowels. It is only the vowels that can have the tonal accents. The alphabets are:

A B D E Ẹ F G GB H I J K L M N O Ọ
a b d e ẹ f g gb h i j k l m n o ọ

P R S Ṣ T U W Y
p r s ṣ t u w y

OCR for different languages has been developed and in use with different application. It will be of great benefit if it is also developed for Yoruba language and for the indigenous people due to its diverse applications such as e-book reader, document editing, speech recognition, text-to-speech among others.

2.0 RELATED WORK

The main function of the OCR process is to scan document in other to extract the text within so it can be easily searched, copied, edited, and matched ([Geo15]). In OCR “1” is black (foreground) and “0” is white (background). It is a step by step process that involves the data acquisition phase which is to digitize the text image, preprocessing, segmentation, feature extraction, classification and recognition of text. Figure 1 shows the block diagram of the OCR.

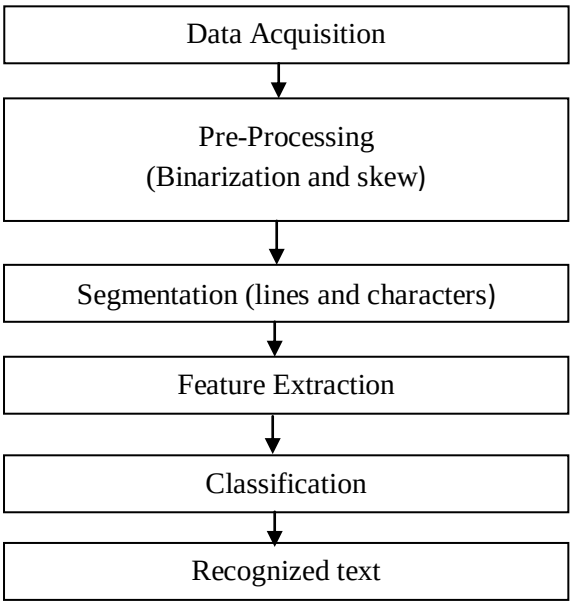


Figure 1: OCR Block diagram ([M+14])

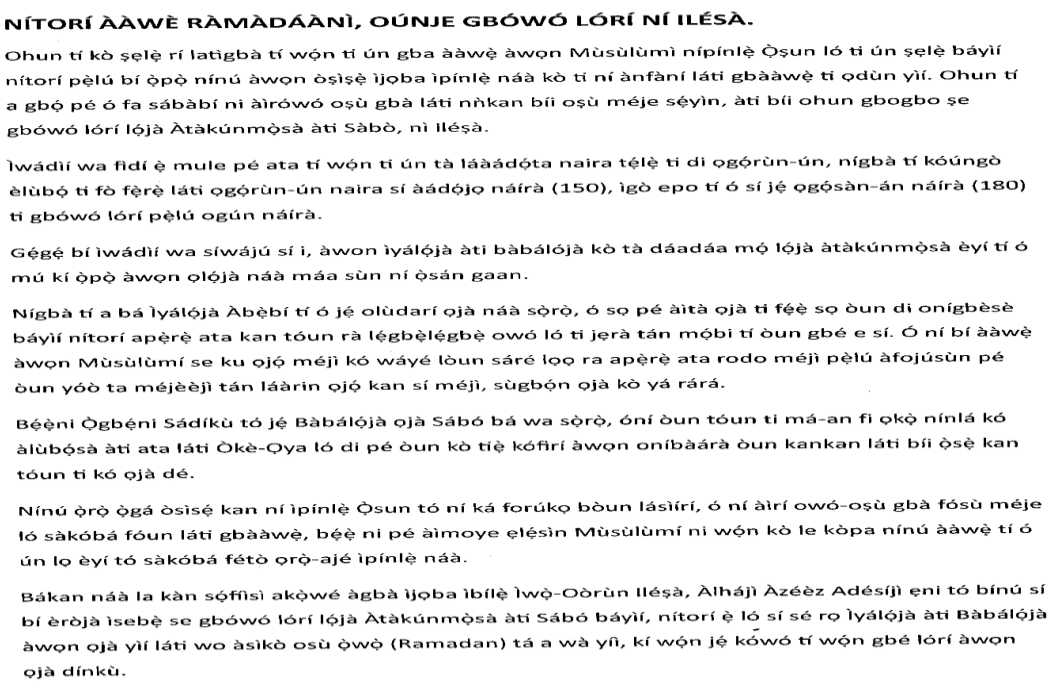
The data acquisition stage is the process of capturing analog materials of characters (paper documents) into digitally recognizable resources. The process is done using an analog-to-digital converter (scanner). The data acquisition stage is followed by the preprocessing stage which involves both binarization and skew. Binarization is used to increase processing speed and reduce the storage require ([CK14]). Skew operation is for proper alignment of scanned document. Segmentation process seeks to breakdown an image of sequence of characters into individual symbols of sub images. Segmentation includes character, word and line segmentation. The character segmentation determines the effectiveness of conventional character recognition systems. ([CK14]). Feature extraction process extracts information from object(s) that are used for classification. The features from isolated characters are extracted which is in form of matrix. The matrix is then used for classification and recognition of characters Patil and Mane ([PM13]) reviewed works that considered the font styles, size, and position using image insight in typed and handwritten recognition of English characters. It was implemented using Matlab and the results were considered to be good with accuracy rate of 96.5%. Tabassam, Syed, Habib & Anoshia ([T+10]) also proposed a system for offline character recognition for isolated characters of Urdu language. The system was implemented using pattern matching with 89% accuracy and 15 characters per second recognition rate. Ajao, Olabiyisi, Omidiora, & Odejobi ([A+15]) evaluated quality of Yoruba handwritten characters using the preprocessing stages of the feature extraction. In the paper, word samples were collected from Yoruba writers and it was concluded that perturbations affected the information on the samples that needed to be removed.

3.0 METHODOLOGY

Optical Character Recognition

Step 1: Scan document in Yoruba and save the scanned image as .bmp file;
Step 2: Read the scanned image file in .bmp format and extract the embedded image byte stream;
Step 3: Convert the colour image from RGB to gray scal;
Step 4: Smooth the gray image using filter;
Step 5: Binarise the smoothed grey image to form edge image;
Step 6: Perform de-skew operation on the scanned image;
Step 7: Perform character row (line) segmentation;
Step 8: Perform character word segmentation;
Step 9: Perform character segmentation;
Step 10: Evaluation

The document scanning involved the use of scanner to convert the document from analog to digital. The digitize document is save in .bmp format which is followed by conversion of coloured image to gray scale image. Binarization is the conversion of grayscale image (0 to 255 pixel values) by selecting a threshold value in between 0 to 255 to binary image (0 and 1 pixel values). Skew operation is used to put the image in right shape. This is to ensure the document is properly aligned. Segmentation derived character or word from the image (input document). Figure 2 shows a .bmp Yoruba scanned text document as a sample input to the proposed model.



NÍTORÍ ÀÀWÈ RÀMÀDÁÀNÌ, OÚNJE GBÓWÓ LÓRÍ NÍ ILÉSÀ.

Ohun tí kò ṣelẹ rí latìgbà tí wọn tí ún gba ààwẹ àwọn Mùsùlùmì nípínlẹ Ọṣun ló ti ún ṣelẹ báyìí nítorí pèlú bí ọpọ nínú àwọn òṣìṣẹ ìjọba ìpínlẹ náà kò ti ní ànfàní láti gbàáwẹ ti ọdùn yìí. Ohun tí a gbọ pé ó fa sábàbí ni àìrówó oṣù gbà láti nǹkan bíi oṣù méje sẹ́yìn, àti bíi ohun gbogbo ṣe gbówó lórí lójà Àtàkúnmọsà àti Sàbò, nì Ilẹṣà.

Ìwádìí wa fidí ẹ mule pé ata tí wọn ti ún tà láàádọta naira tẹlẹ ti di ogọrùn-ún, nígbà tí kóúngò èlùbọ ti fò fẹrẹ láti ogọrùn-ún naira sí àádọjọ náírà (150), ìgò epo tí ó sí jẹ ogósàn-án náírà (180) ti gbówó lórí pèlú ogún náírà.

Gẹgẹ bí ìwádìí wa síwájú sí i, àwon ìyálójà àti bàbálójà kò tà dáadáa mọ lójà àtàkúnmọsà èyí tí ó mú kí ọpọ àwọn olójà náà máa sùn ní ọsán gaan.

Nígbà tí a bá ìyálójà Àbẹbí tí ó jẹ olùdarí ojà náà sọrọ, ó so pé àìtà ojà ti fẹẹ so òun di onígbèsè báyìí nítorí apèrè ata kan tóun rà lẹgbẹlẹgbẹ owó ló ti jerà tán mọbi tí òun gbé e sí. Ó ní bí ààwẹ àwọn Mùsùlùmí se ku ojó méjì kó wáyé lòun sáré loọ ra apèrè ata rodo méjì pèlú àfojúsùn pé òun yóò ta méjèèjì tán láàrin ojó kan sí méjì, sùgbọn ojà kò yá rárá.

Bẹẹni Ọgbẹni Sádíkù tó jẹ Bàbálójà ojà Sábó bá wa sọrọ, óní òun tóun ti má-an fi okọ ninlá kó àlùbósà àti ata láti Òkè-Oya ló di pé òun kò tiẹ kóñrí àwọn oníbàárà òun kankan láti bíi ọsẹ kan tóun ti kó ojà dé.

Nínú ọrọ ọgá òsìsẹ kan ní ìpínlẹ Ọsun tó ní ká forúkọ bòun lásìírí, ó ní àìrí owó-oṣù gbà fósù méje ló sàkóbá fóun láti gbàáwẹ, bẹẹ ni pé àìmoye elẹsìn Mùsùlùmí ni wọn kò le kòpa nínú ààwẹ tí ó ún lọ èyí tó sàkóbá fétò ọrọ-ajé ìpínlẹ náà.

Bákan náà la kàn sọ̀fìsì akọ̀wé àgbà ìjọba ìbílẹ ìwọ-Oòrùn Ilẹṣà, Àlhájì Àzéèz Adésíjì ẹni tó bínú sí bí èròjà ìsebẹ se gbówó lórí lójà Àtàkúnmọsà àti Sábó báyìí, nítorí ẹ ló sí sé ro ìyálójà àti Bàbálójà àwọn ojà yìí láti wo àsìkò osù ọ̀wọ̀ (Ramadan) tá a wà yíì, kí wọn jẹ kówó tí wọn gbé lórí àwọn ojà dínkù.

Figure 2: Scanned Image (Yoruba Text)

The metric used for the evaluation of the performance of the proposed Yoruba OCR model are Precision and Recall. They are robust evaluation metrics that have been extensively used in information retrieval studies.

$$\text{Precision} = \frac{|S \cap I|}{|S|} \quad (1)$$

$$\text{Recall} = \frac{|S \cap I|}{|I|} \quad (2)$$

Where:
I is the sample you give to your system
S is what your system produced from the sample (system output)
$|\ |$ implies the cardinality of the set i.e., the number of characters in the set. So
$|I \cap S|$ is the number of characters in the resultant set of I intersect S
$|I|$ is the number of characters in the set I
$|S|$ is the number of characters in the set S

4.0 RESULT AND DISCUSSION

4.1 Result

The input/output of the proposed OCR model for Yoruba are as presented in the following subsections. The result of the performance evaluation of the model is also presented.

4.1.1 The OCR Input

The OCR input is a sample document which is the .bmp input file for processing Yoruba scanned text document as shown uploaded in figure 3. The document used is a Yoruba character document typed and scanned. The document was printed on a plain white paper and scanned with a scanning machine.

4.1.2 Processing Steps

The tasks in the processing steps of OCR includes Grayscaling, Smoothing, Edge detection, Skewing and Segmentation are shown in figure 4. The result of each of the tasks is presented as follows.
Figure 5 presents the result of the Grayscaling step. The original input is in the left hand side and the grayscal output is in the right hand side. Similarly, the original document input and the output results for the smoothing, edge detection, skewing and segmentation are presented in figures 6, 7, (8 & 9), and 10 respectively.
Segmentation: Line and character segmentation was performed on the image. Figure 10 shows the result of line segmentation. It identifies each line of characters, taking in to cognizance the ascent of each character, separating each character from the next.

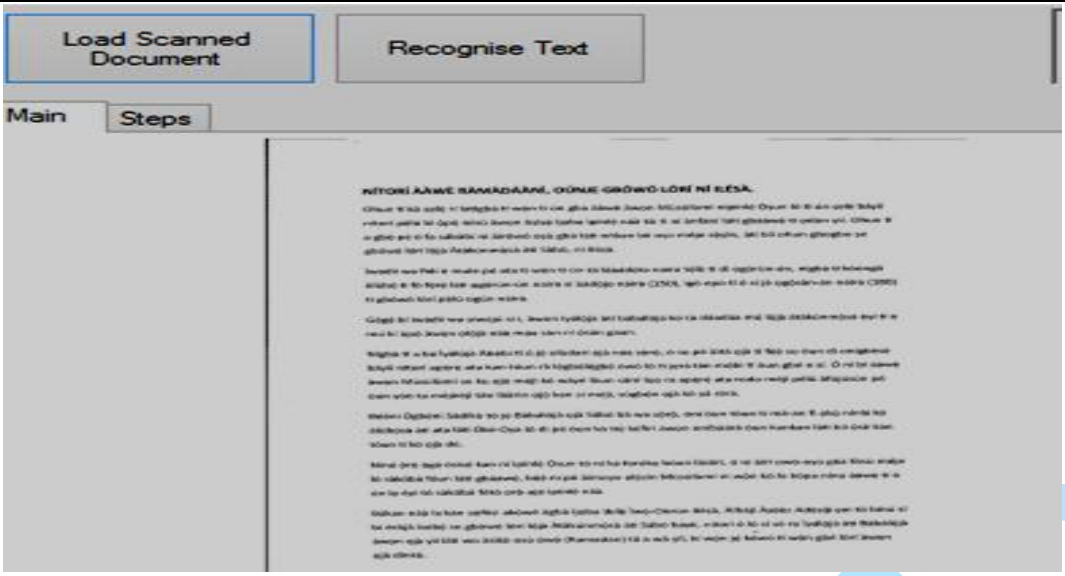


Figure 3: .bmp Yoruba character document

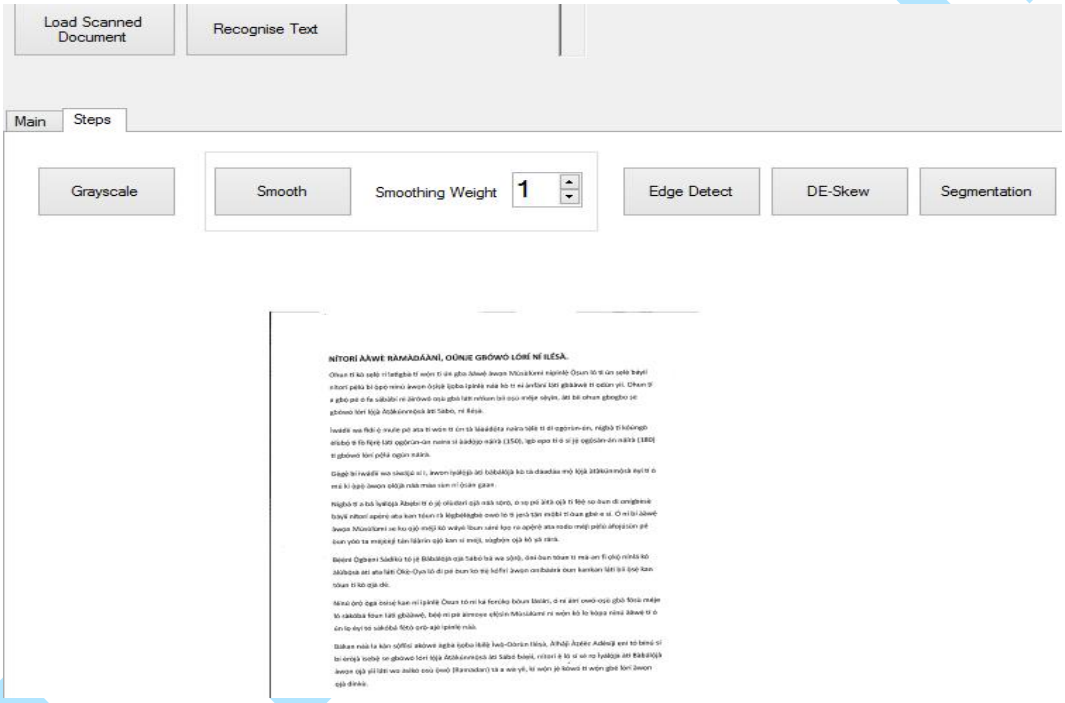


Figure 4: OCR Steps

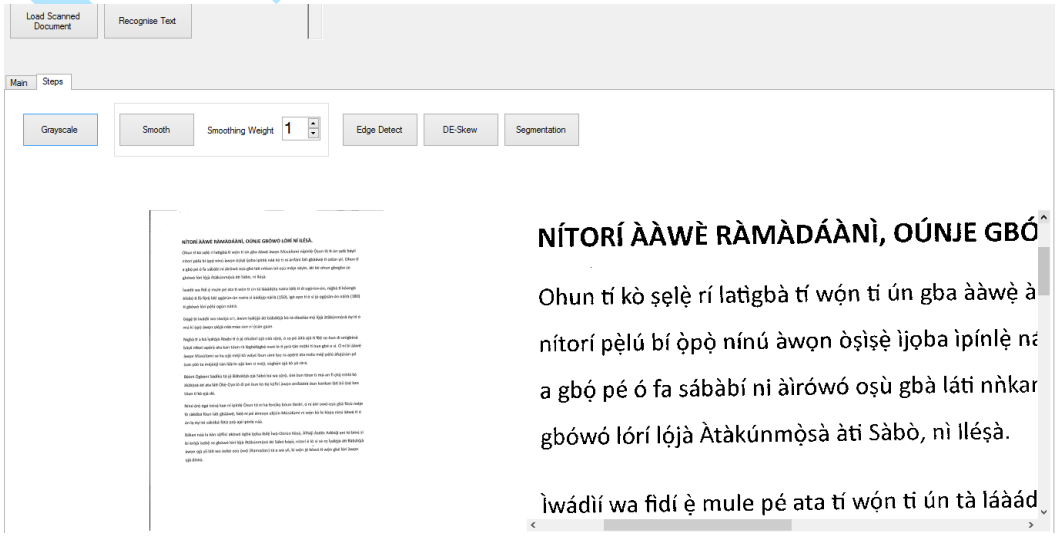


Figure 5: Grayscale implementation

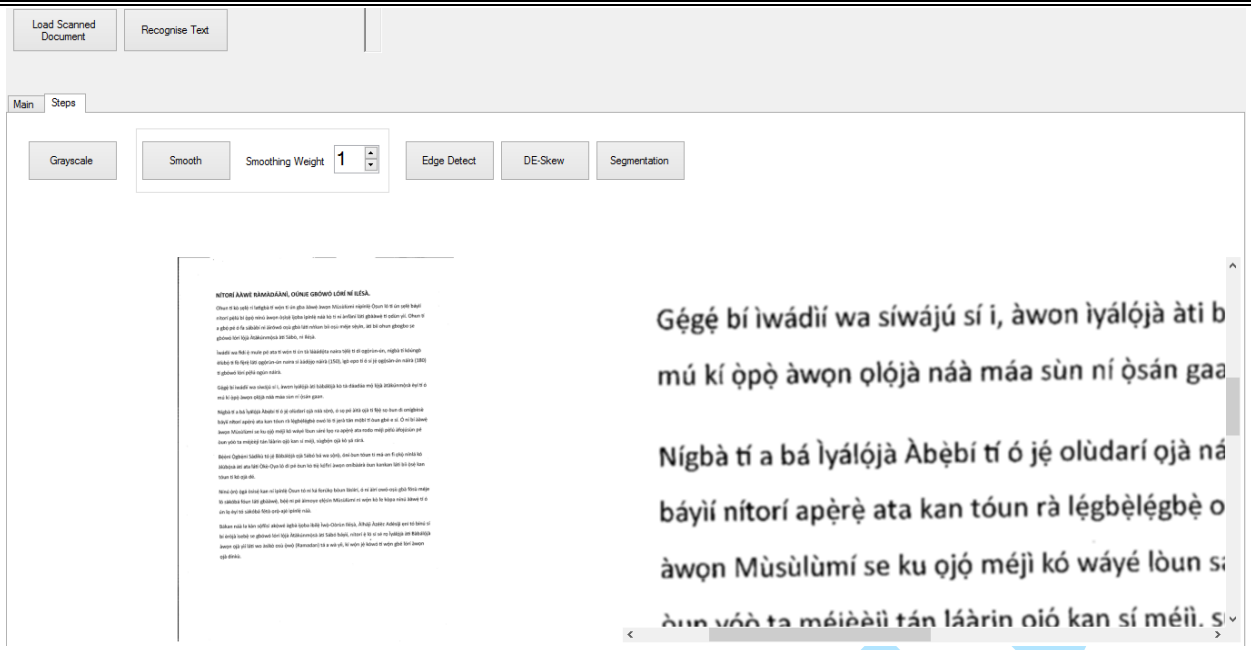


Figure 6: Smoothening Implementation

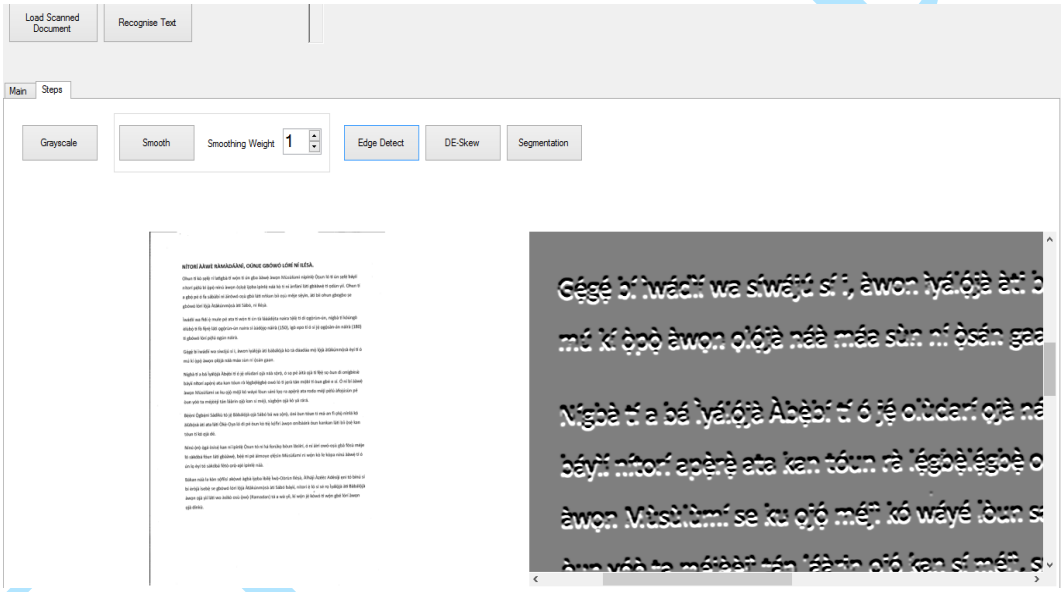


Figure 7: Edge Detection Implementation

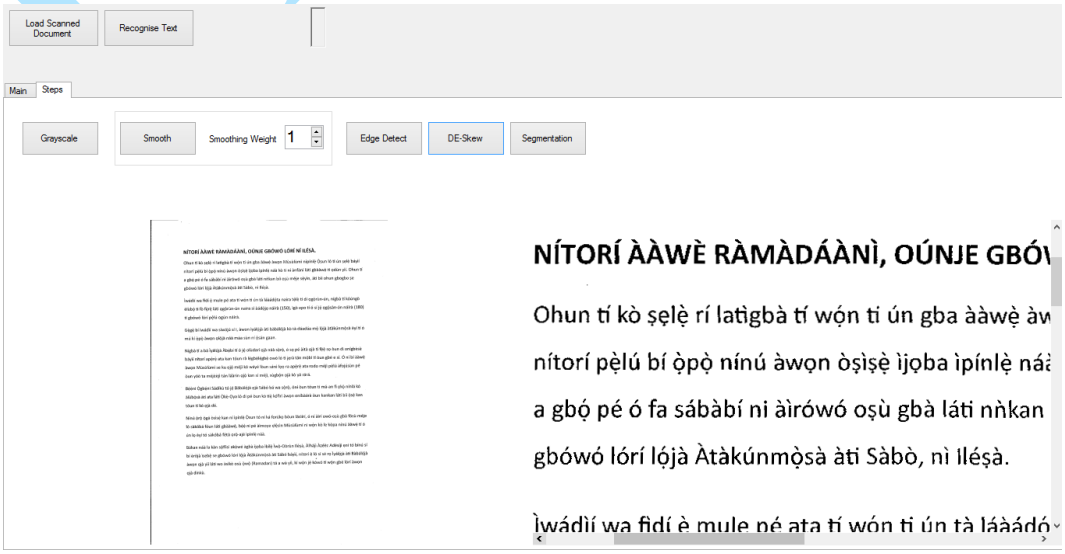


Figure 8: De-Skew Implementation

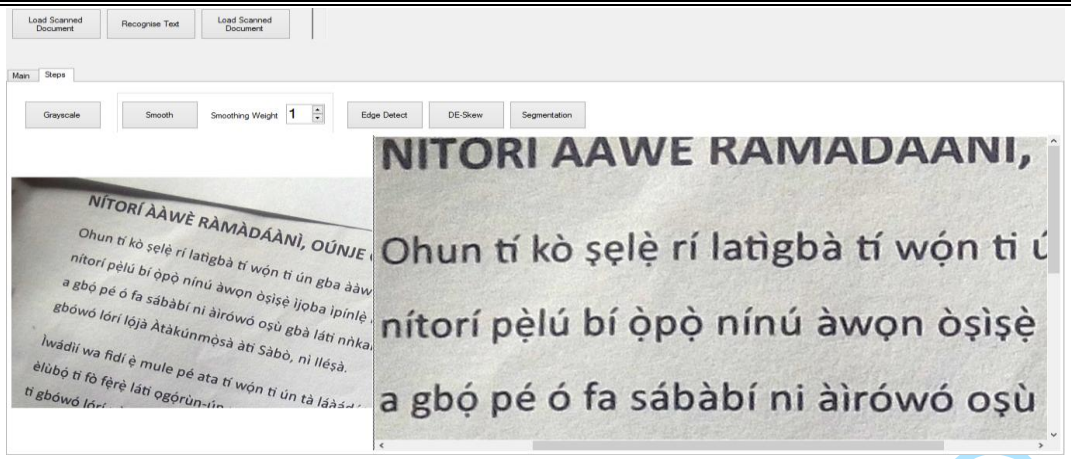


Figure 9: De-Skew Implementation

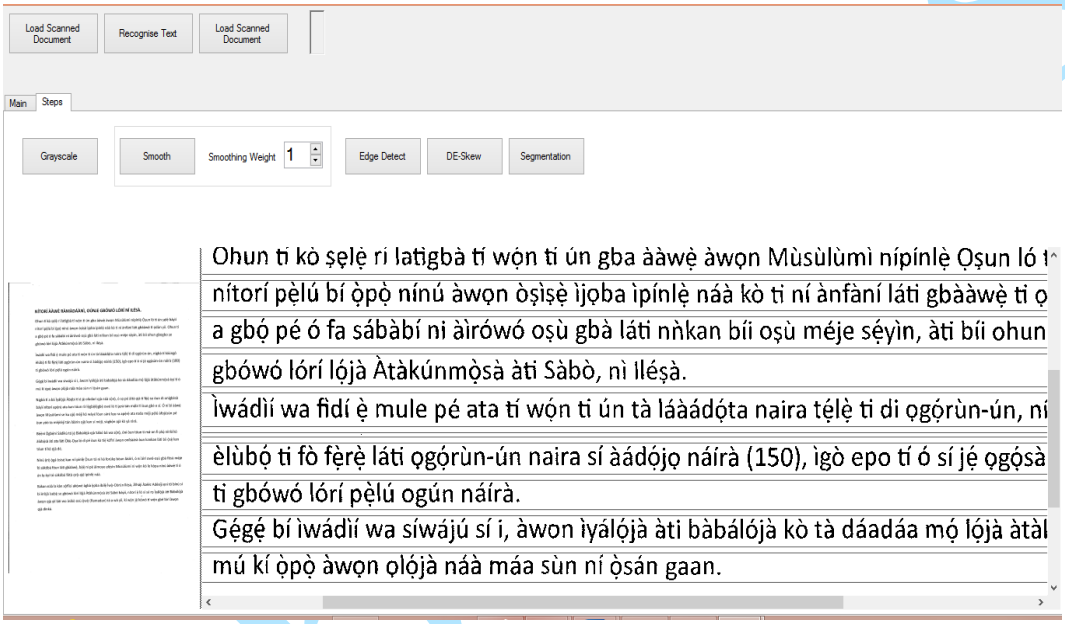


Figure 10: Segmentation of Yoruba text image

4.1.3 OCR Model Output

Figure 11 shows the output of the proposed OCR model. The input data is on the left hand side of the image, while the output with system recognized characters on the right hand side. The output is further used for the performance evaluation.

4.1.4 Character Recognition Accuracy

An error is said to occur if there is substitution or insertion required to correct the generated text. The information retrieval metrics Recall and Precision are used to measure the performance of the proposed Yoruba OCR model. As earlier stated, precision is the fraction of retrieved text that are relevant and Recall is the fraction of relevant text that are retrieved and they are determined using the equation (1) and (2) in section 3.

In the evaluation, 6 sample pages were used. Figures 12 to 17 presents these sample pages.

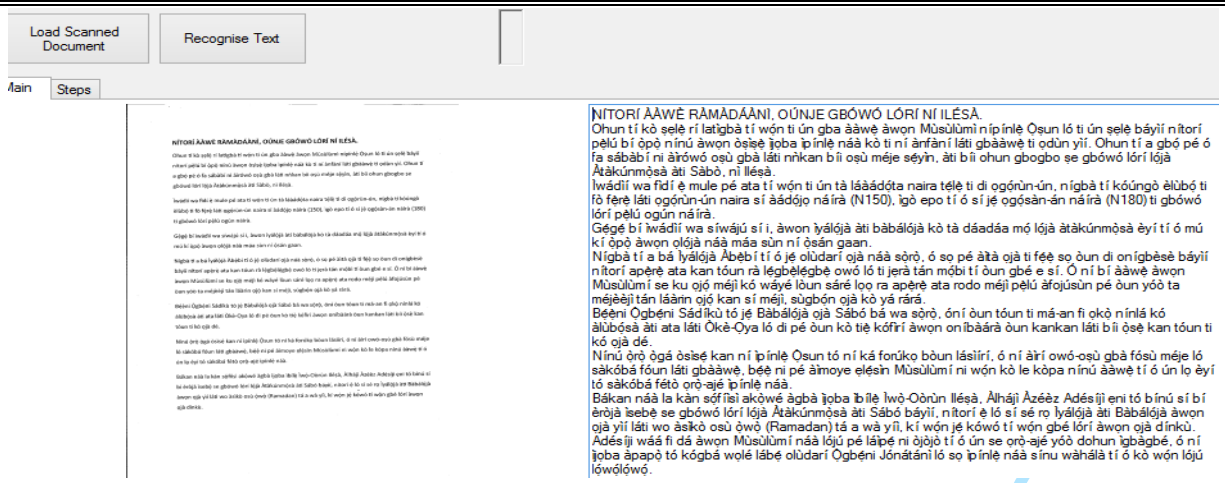


Figure 11: Output

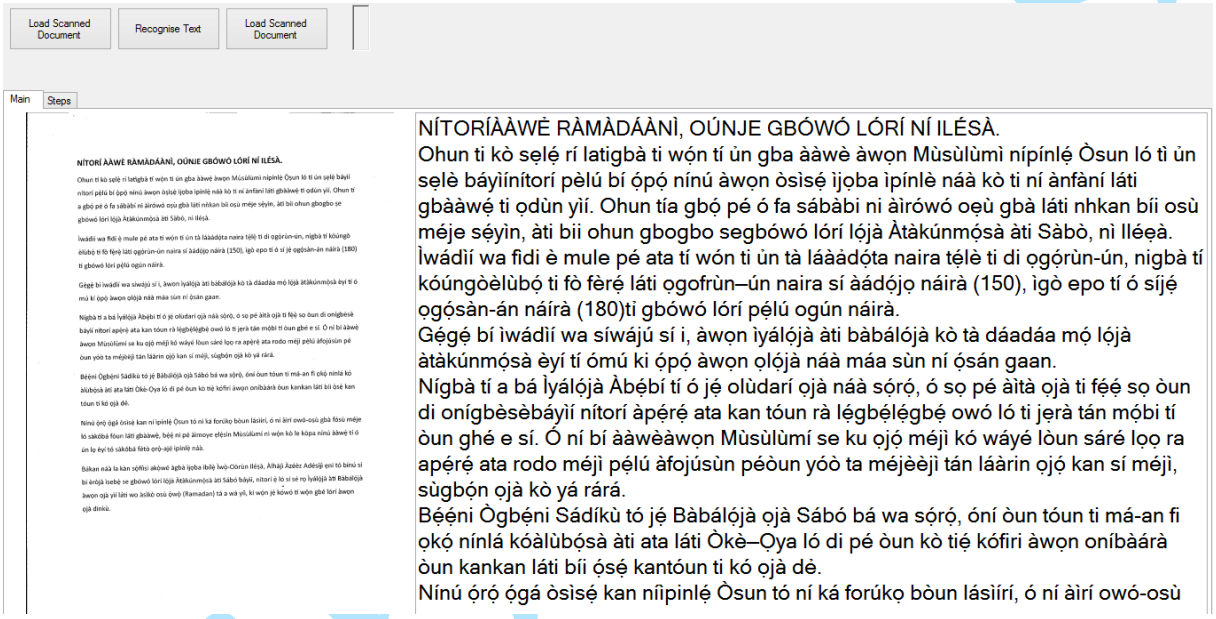


Figure 12: Sample Page 1

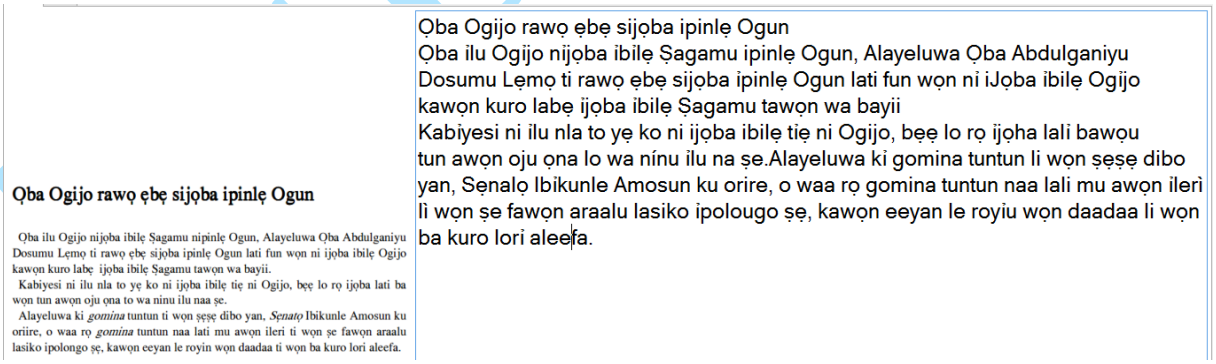


Figure 13: Sample Page 2

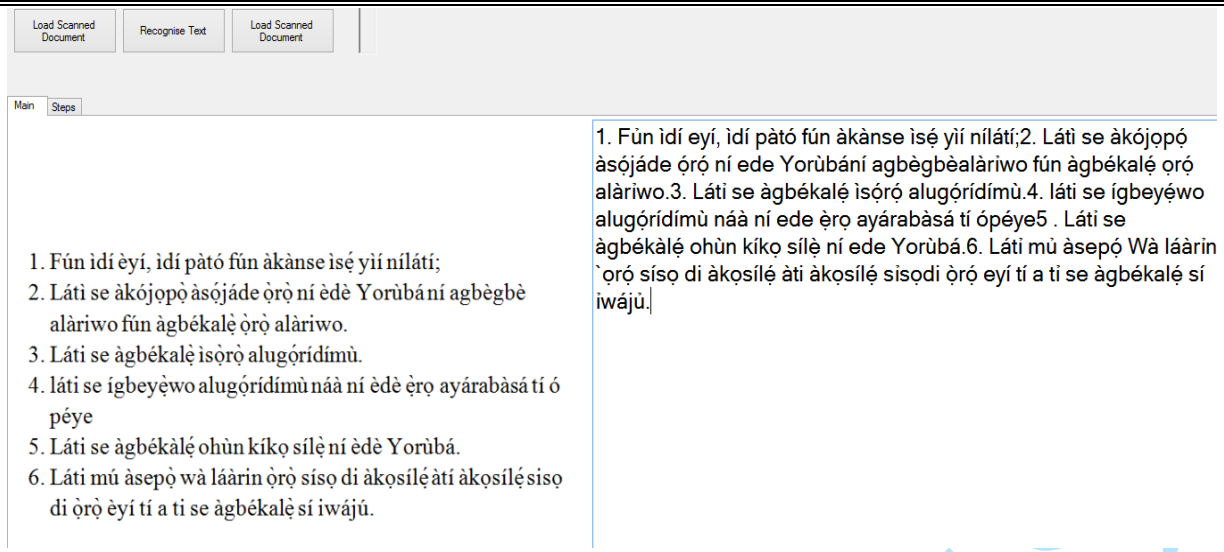


Figure 14: Sample Page 3

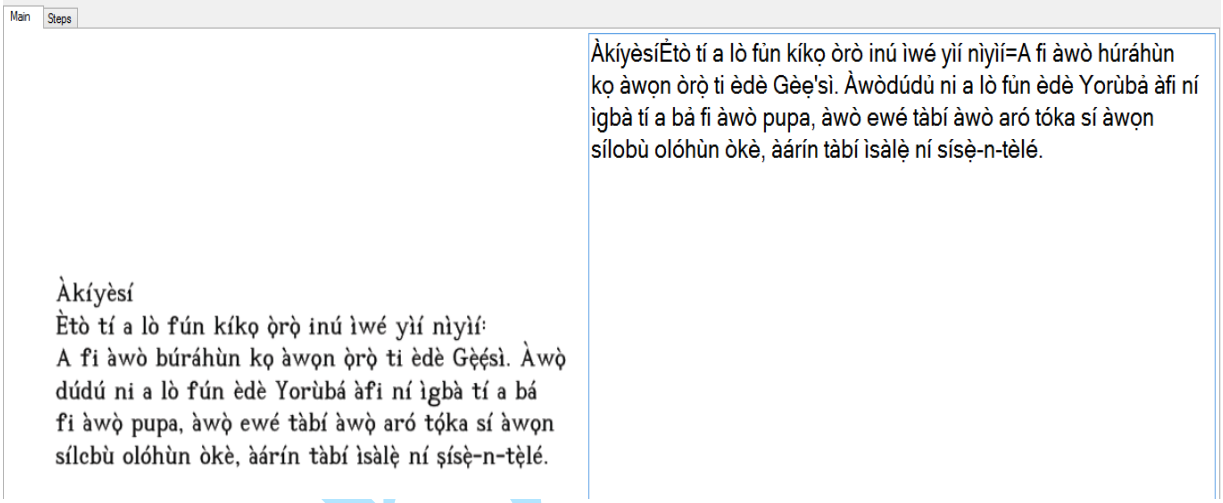


Figure 15: Sample Page 4

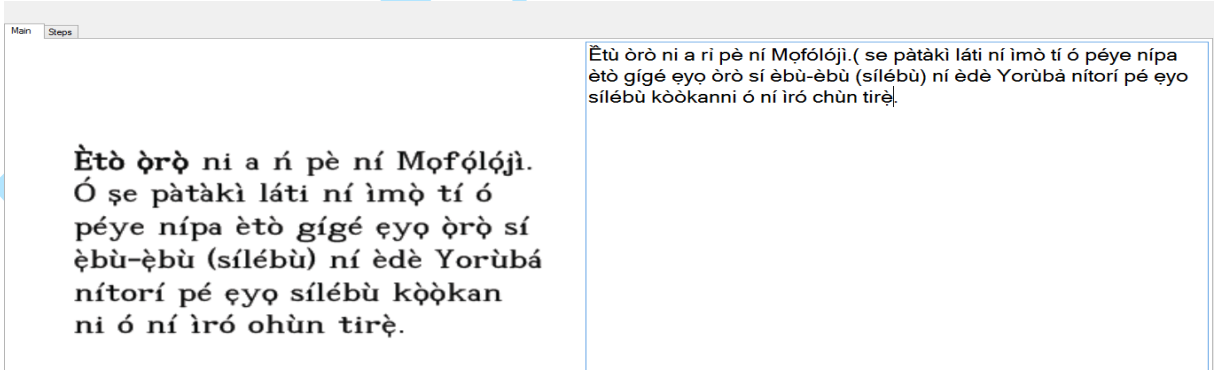


Figure 16: Sample Page 5

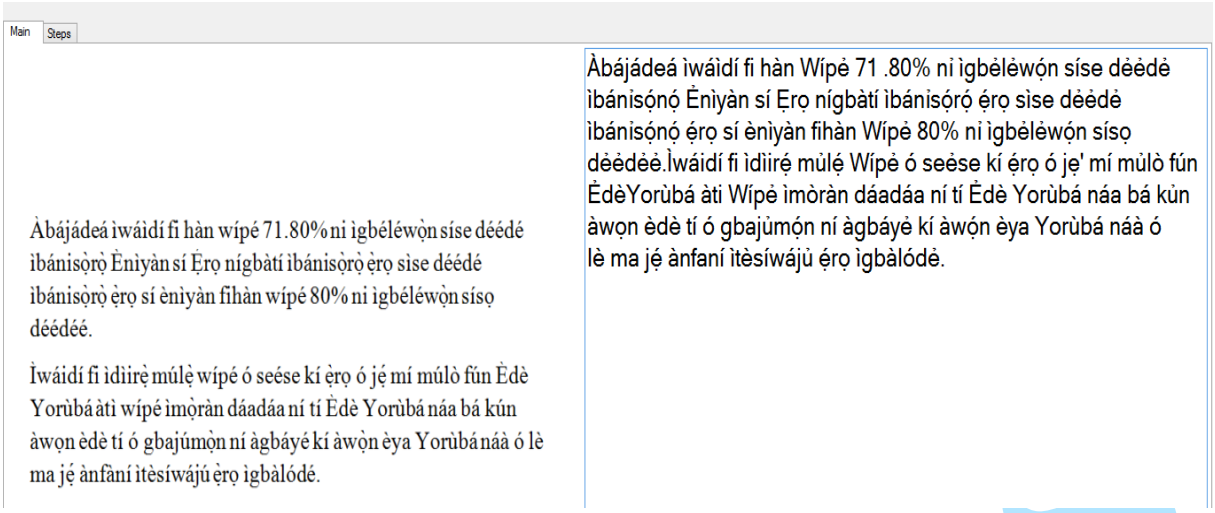


Figure 17: Sample Page 6

Table 1 shows the precision and recall results for the 6 pages samples. The total number of characters used is 4665 out of which 4540 characters are correctly recognized.

Table 1: Character accuracy for 6 pages samples

Sample	Relevant Characters	Retrieved Characters	Precision %	Recall %
Page 1	2495	2496	99.96	100
Page 2	812	823	98.66	100
Page 3	308	406	75.86	100
Page 4	283	290	97.59	100
Page 5	183	186	98.39	100
Page 6	459	464	98.92	100

Table 2 shows the correct character and the generated character. Generally, substitution occurred with one consonant lower case ṣ that is substituted with s or space or s_ or ẹ and lowercase vowel ọ is sometimes substituted with ẹ or ọ for sample 1 in figure 12. Sample 1also included errors with characters b, t, and u due to noise on page (scanned document) as shown in the output in Table 2.

Table 2: Correct character with corresponding generated character

Correct	Generated
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
A	A
B	B
D	D
E	E
Ẹ	Ẹ
F	F
G	G
I	I
H	H
J	J
K	K
L	L
M	M
N	N
O	O
Ọ	Ọ
P	P
R	R
S	S
Ṣ	Ṣ
T	T
U	U
W	W
Y	Y
A	A
B	b, h,
D	D
E	e
Ẹ	ẹ
f	f
g	g
i	i
h	h
j	j
k	k
l	l
m	m
n	n
o	o
ọ	ọ
p	P
r	R
s	S
ṣ	s, _,s_
t	t , l
u	u, n
w	w W
y	Y
À	À
à	À
Á	Á
á	Á
È	È
È	è
É	É
É	é
Ẹ	Ẹ
Ẹ	ẹ
È	È
Ẹ	è
Ì	Ì
Ì	ì
Í	Í
Í	í
Ò	Ò
Ò	ò
Ó	Ó
Ó	ó
Ọ	Ọ
Ọ	ọ
Ọ	Ọ
Ó	ó
Ù	Ù
Ù	ù
.	.
,	,

4.2 Discussion

The proposed Yoruba OCR model has shown a significant performance result in the recognition of Yoruba characters. It showed a 100% recall which implies that all the input characters are accurately recognized by the OCR model. The precision of the model varies very slightly between 97% to 100% except in the sample page 3 where a precision of approximately 76% is observed. Precision is the number of relevant characters among the recognized characters. This lower precision is as a result of fonts used for the characters.

CONCLUSION

The Yoruba OCR model encompassed various stages that include gray scal, smoothing, skew, segmentation and recognition. The model created can be used to digitize Yoruba characters. This can be used in speech recognition, text-to-speech and also to digitize old hardcopy text that would have been redundant after a period of time. The model was evaluated using recall and precision metrics with the 100% performance in recall while precision varies between 76%, 97%, and 100% accuracy. In future, the model can be extended to other Nigerian languages like Hausa and Igbo.

REFERENCES

[A+14] **Abiola O., Adetunmbi A, Fasiku A, Olatunji K.** - *Web-based English to Yoruba noun-phrases machine translation system.* International Journal of English and Literature. 5(3), 71-78, 2014.

[A+15] **Ajao F., Olabiyisi O., Omidiora O., Odejobi O.** - *Yoruba Handwriting Word Recognition Quality Evaluation of Preprocessing Attributes using Information Theory Approach.* International Journal of Applied Information Systems (IJAIS). 9(1) 18-23, 2015.

[CK14] **Chandarana J., Kapadia M.** - *Optical Character Recognition.* International Journal of Emerging Technology and Advanced Engineering. 4(5) 219-223, 2014.

[Geo15] **George N.** - *Document processing applications.* Rensselaer Polytechnic Institute, NY,USA, 2015.

[Kar08] **Karry F.** - *Human computer interaction: Overview on state of the art.* International Journal on Smart Sensing and Intelligent Systems, 1, 137-159, 2008.

[M+14] **Mohammad F., Anarase, J., Shingote M., Ghanwa P.** - *Optical Character Recognition Implementation Using Pattern Matching.* International Journal of Computer Science and Information Technologies. 5(2). 2088-2090, 2014.

[PM13] **Patil J., Mane A.** - *Multi Font and Size Optical Character Recognition Using Template Matching.* International Journal of Emerging Technology and Advanced Engineering. 3(1). 504-506, 2013.

[T+10] **Tabassam N., Syed, A., Habib R., Anoshia F.** - *Optical Character Recognition System for Urdu (Naskh Font) Using Pattern Matching Technique.* International Journal of Image Processing (IJIP). 3 (3) 92-104, 2010.