TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 13, No. 1, January 2015, pp. 166 ~ 173
DOI: 10.11591/telkomnika.v13i1.6782

Yoruba Language and Numerals' Offline Interpreter Using Morphological and Template Matching

Olakanmi O. Oladayo
Electrical and Electronic Engineering,
Technology Drive, Office 6, New Faculty of Engineering Building.
University of Ibadan, Ibadan Nigeria
email: olakanmi.oladayo@ui.edu.ng

Abstract

Yoruba as a language has passed through generation reformations making some of the old documents in the archive to be unreadable by the present readers. Apart from this, some Yoruba writers usually mixed English numerals while writing due to brevity and conciseness of English numerals compare to Yoruba numerals which are combination of several characters. Re-typing such historical documents may be time consuming, therefore a need for an efficient Optical Character Reader (OCR) which will not only effectively recognize Yoruba texts but also converts all the English numerals in the document to Yoruba numerals.Several Optical Character Reader (OCR) systems had been developed to recognize characters or texts of some languages such as English, Arabic, Japanese, Chinese, and Korean, however, despite the significant contribution of Yoruba language to historical documentation and communication, it was observed that there is no particular OCR system for the language. In this paper correlation and template matching techniques were used to develop an OCR for the recognition of Yoruba based texts and convert English numerals in the document to Yoruba numerals. Experimental results show the relatively high accuracy of the developed OCR when it was tested on all size Yoruba alphabets and numerals.

Keywords*: OCR, Yoruba, pattern recognition, image, template matching*

1. Introduction

Yoruba is the most documented West African language. Yoruba is spoken by 18,850,000 people in Nigeria. The total population of native speakers in all countries is about 20,000,000. The number rises to 22,000,000 if we also include second-language speakers. The language has numerous dialects spoken in different areas of Nigeria. Within Nigeria the language is spoken in the areas of Oyo, Ogun, Ondo Osun, Kwara, Lagos and the western part of Kogi State. It is also spoken in Benin, Togo, and by immigrants in the United Kingdom and the USA. Yoruba is one of the 12 Edekiri languages of the Yoruboid group that also includes Igala. The Yoruboid group belongs to the Defoid languages of the Benue-Congo group and ultimately to the Volta-Congo, and Atlantic-Congo groups of the Niger-Congo Family of 1419 languages mostly spoken in Central and South Africa [6].

Image recognition is the process of identifying and detecting an object or a feature in a digital image or video. This concept is used in many applications like systems for factory automation, toll booth monitoring, and security surveillance. Typical image recognition algorithms include:

1) Optical character recognition
2) Pattern and gradient matching
3) Face recognition
4) License plate matching
5) Scene change detection

It has become a trend to document most of the documents in the archives using scanner, however, these documents cannot be edited or read thereafter by computer systems. Due to the fact that scanner scans documents as an image not as encoded set of characters. Optical Character Reader (OCR) system does electronic translation of handwritten or printed text into machine encoded text. OCR is widely used to convert books and documents into electronic files and to computerize a record keeping system in an office. OCR makes it possible

Received August 2, 2014; Revised September 18, 2014; Accepted October 16, 2014

to edit such document, search for a word or phrase, store it more compactly, display or print a copy and apply techniques such as machine translation, text-to-speech and text mining to it.

Optical Character Recognition study was started by Tyurin a Russian scientist [1]. The first modern character recognizers appeared in the middle of the 1940s with the development of the digital computer. The early work on the automatic recognition of characters has been concentrated either upon well printed text or upon small set of well distinguished handwritten text or symbols, although, successful but had been implemented mostly for Latin characters and numerals. Besides some studies on Japanese, Chinese, Hebrew, Indian and Arabic charades and numerals in both printed and handwritten cases were also considered by some OCR systems. The developments in OCR until 1980s suffered from lack of advanced algorithm, powerful computing hardware and optical devices. With the outward explosion on the computing technology development, the previously proposed methodologies found a fertile environment for rapid growth in many application areas. Presently, renewed vigours are being put in the optical character recognition research. One of these is recognition of printed and handwritten documents. More sophisticated algorithms which utilize advanced methodologies are being developed.

In this work two methodologies are combined to achieve an efficient Yoruba OCR system which will be able to recognize off-line typed and handwritten Yoruba documents and convert English numerals to Yoruba numerals. The remaining part of this paper is arranged as follows: section 2 is the review of related works on OCR systems and methodologies. The design methodology and working principle of the system are explained in section 3. Section 4 contains the test results and conclusion.

2. Related Works

Reference [8] described a complete system for the recognition of unconstrained handwritten Arabic words using over-segmentation of characters and variable duration hidden Markov model (VDHMM). In this, a segmentation algorithm was used to translate the 2-D image into 1-D sequence of sub-character symbols. This sequence of symbols was modeled by the VDHMM. The shape information of character and sub-character symbols was compactly represented by forty-five features in the feature space. The feature vector was modeled as an independently distributed multivariate discrete distribution. And the variable duration state is used to resolve the segmentation ambiguity among the consecutive characters.

Different methodologies on how the quality of the captured camera image could be improved had been thoroughly considered by various researches. For example, reference [2] analyzed the quality of such captured image for optical character recognition. In their work different means of improving transcription and recognition was proposed. Also, reference [18] proposed a new perspective rectification system based on vanishing point detection. Their system achieved both the desired efficiency and accuracy using a multi-stage strategy: at the first stage, document boundaries and straight lines are used to compute vanishing points; at the second stage, text baselines and block aligns are utilized; and at the last stage, character tilt orientations are voted for the vertical vanishing point. A profit function was introduced to evaluate the reliability of detected vanishing points at each stage. If vanishing points at one stage are reliable, then rectification is ended at that stage. Otherwise, multi-stage strategy method continues to obtain more reliable vanishing points in the next stage.

Research has shown that Character degradation affects machine printed character recognition. Two main reasons for degradation were extrinsic image degradation such as blurring and low image dimension, and intrinsic degradation caused by font variations. A recognition method that combines two complementary classifiers is proposed in reference [17]. The local feature based classifier extracts the local contour direction changes, which is effective for character patterns with less structure deterioration. The global feature based classifier extracts the texture distribution of the character image, which is effective when the character structure is hard to discriminate. The two complementary classifiers are combined by candidate fusion in a coarse-to-fine style. Experiments are carried on degraded Chinese character recognition.

Reference [13] worked on Character recognition system Telugu; one of the ancient languages of South India. It has a complex orthography with a large number of distinct character shapes composed of simple and compound characters. In this work, structural features of the

syllable and the component model were combined to extract middle zone components. The shape of the middle zone components is closely related to a circle whereas other components are found with different topological features.

A simple and effective template matching method for identification of Musnad characters was introduced in reference [10[. The characters were extracted from input image and normalized. During recognition, the extracted character was compared to each template in the database to find the closest representation of the input character. The matching metric was computed using 2-D correlation coefficients approach to identify similar patterns between the test image and the database images.

In reference [5], a novel approach to efficiently recognize handwritten numerals was proposed. This approach exploits a two-stage framework by using difference features. In the first stage, a regular SVM is trained on all the training data; in the second stage, only the samples misclassified in the first stage are specially considered. The number of misclassifications is often small because of the good performance of SVM. This will present difficulties in training an accurate SVM engine only for these misclassified samples. We then further propose a multi-way to binary approach using difference features. This approach successfully transforms multi-category classification to binary classification and expands the training samples greatly.

2.1. Overview of Yoruba Orhography

In its written form, Yoruba uses the Roman alphabet. It has 25 letters as shown in fig. 2. The letter 'p' is always pronounced as 'k' and 'p' combined. Yoruba orthography does not use the letters c, q, v, x, z. Yoruba has three basic tones, high, mid, and low, which are indicated in the orthography. The high is marked with an acute accent (e.g. á), the low with a grave accent (à), and the mid tone usually left unmarked. These marks are usually placed on the vowels. In some circumstances the mid tone is indicated with a 'macron'. The language has been written since the early 19th century, although there have been many changes in aspects of its orthographic representation. In the 1960s, the then Ministry of Education within the Western Region of Nigeria, which was where most of the Yoruba speaking community is located, formed two committees to consider a standard orthography for the language. The more influential of these two, the Yoruba Orthography Committee was set up in 1966. The report which this second orthography committee submitted in 1966 became the basis for the creation and introduction into schools of the standard Yoruba orthography [6].

Table 1. English numerals and their equivalent Yoruba numerals

English	*1*	*2*	*3*	*4*	*5*	
Yoruba	Eni	Eji	Eta	Erin	Arun	
English	6	7	8	9	10	0
Yoruba	Efa	Eje	Ejo	Esan	Ewa	Odo

.

Table 2. Yoruba upper and lower alphabets

A a	B b	D d	E e	Ẹ ẹ
F f	G g	GB gb	H h	I i
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

2.2. Yoruba OCR System Methodology

OCR as earlier stated is the science that entails the description or classification of character measurements that usually based on some models. OCR is one of the categories of image recognition. There is various character recognition methods used in developing character recognizer. These methods are: neural network, moment based approach, contour based approach, template matching and morphological approach. In this work template matching and morphological techniques are used to recognize Yoruba texts.Template matching refers to the process of detecting an object having a certain size, shape and orientation in an image by applying an operator containing positive weights in a region resembling the objects to be

detected and containing negative weights in a region surrounding the positive weight [15]. Morphology as derived from biology is a branch of biology which deals with the form and animals and plants. It is adopted in this context as a tool for extracting image components that are useful in the representation and description of the region shape. There are several procedural steps engaged in achieving morphological techniques. These include filtering, thinning, pruning, erosion and dilation, opening and closing.

3. Yoruba OCR Implementation using Template Matching and Morphological Technique

Template matching and morphological techniques as stated earlier, are OCR recognition techniques. These algorithms involve features extraction and classifier. In template matching image pixels are used as the features being extracted from both the input character and the classified characters. The classifier compares the input character features with a set of character template in the character class. In this context the character class contains numerals, upper and lower cases of Yoruba characters as shown in Figure 1 and Figure 2. The absolute value of the classifier procedure which is the correlation coefficient between the input character and the considered character template is used to morphologically determine the template with a closest correlation match.

Formally,

$$X = (U, L, N, P) \quad (1)$$

$$c_n = \{y \in (U, L)\} \quad (2)$$

Where:
$U\ is\ the\ set\ of\ uppercase\ Yoruba\ characters$
$L\ is\ the\ set\ of\ lowercase\ Yoruba\ characters$
$N\ is\ the\ set\ of\ \ Yoruba\ numbers$
$P\ is\ the\ set\ of\ \ Yoruba\ punctuation\ marks$

The transformation function δ on character c is:

$\tau_n : c_n\ X\ \delta\ \rightarrow\ \tau_n$
$\tau\ is\ the\ set\ of\ \ templates\ of\ characters$

In the character class some of the characters were written in different ways in order to accommodate different ways of writing. The proposed Yoruba OCR system, as shown in figure 3, is grouped into three processing levels which are low level processing, intermediate level and high level processing. These are implemented using 64-bit Matlab version 7.8.0.387 and the input texts are built with paint brush and text.

3.1. Low Level Processing

As shown in the Figure 3, low level processing involves image acquisition and pre-processing of the acquired images. Image acquisition stage acquires image of the document or characters to be recognized. Most time input character image is of finite resolution which ultimately affects the quality of its transformation, therefore, pre-processing becomes necessary. The pre-processing stage includes colour normalization, scaling filtering and thinning. Colour normalization is used to change input character foreground colour to black and background colour to white. To achieve this, histogram technique was used. The input character was used to form histogram of single class which was grouped into intervals. Over each of these intervals a vertical rectangle is drawn with its area proportional to the number of point falling into that interval. The luminance of the image was determined using equation 3. Figure 2a shows input image before normalization while figure 2b and 2c depict the input image after normalization and filtering respectively.

$$Lu = 0.3R + 0.59G + 0.11B \quad (3)$$

Normalization algorithm:

1) Select the relevant part of the character.
2) Determine the threshold for the colour normalization
3) Process the image from top corner line by line
4) Store the R,G,B value of each pixel
5) Determine Lu using equation 1
6) If Lu< threshold value then turn the pixel black otherwise white.
7) Repeat for the whole input image

The image scaling scales the input character image up or down depending on the original size. This was done to reduce the recognition time and error rate as large character images would take longer time to process while small image may be difficult to recognize. After scaling the character becomes blocky and hence the smoothening filtering stage removes the spike edges. This stage also contains smoothening filter, low pass filter. These filters are used to reduce blurring and noise. Also, implemented in the low level processing is the thinning which converts any elongated parts or strips in the image regardless of their bits into narrow strips that are only about one pixel wide.

3.2. Intermediate Level Processing

Intermediate Level Processing (ILP) in the in figure 3 involves image rotation and segmentation. Sometimes input character image may not be properly aligned in angular fashion with respect to the character template set. An instance of this will be corrected by realign the image OCR. Segmentation which forms the core of IL processing stage partitions the input image into its constituent characters. Shown below is the algorithm used for segmentation:

Segmentation algorithm:

1) Scan the image from right to left row wise
2) Add and count all the x coordinates
3) Determine the x-coordinate of the centroid using $xcentroid = \sum(x)/n$ where n is the total number of the centroid.
4) Determine the y-coordinate of the centroid using $ycentroid = \sum(y)/n$ where n is the total number of the centroid.

3.3. Representation and Description

Representation maps the scanned character image to form suitable for subsequent computer processing while description is a feature selection which deals with extracting features in some quantitative manner or differentiating one class of objects from another. This was achieved using internal characteristics, that is, the pixels compromising the region.

3.4. Knowledge Base

The knowledge base contains the numbers, punctuation, upper and lower cases of Yoruba alphabets as shown in Figure 1 and 2. It is basically a database of typed and handwritten English alphabets, numbers, and punctuations. Individual character images in the knowledge base are used to generate the correlation values for the input character image and output character text.

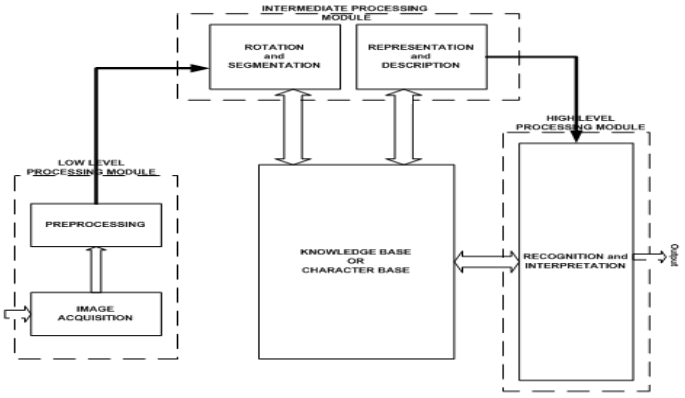

Figure 1. Schematic of the off-line Yoruba Optical Character Reader

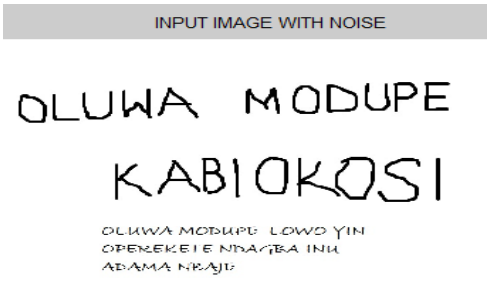

Figure 2(a). Input image character before normalization

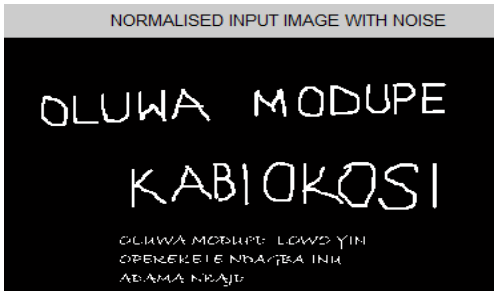

Figure 2(b). Input image text after normalization

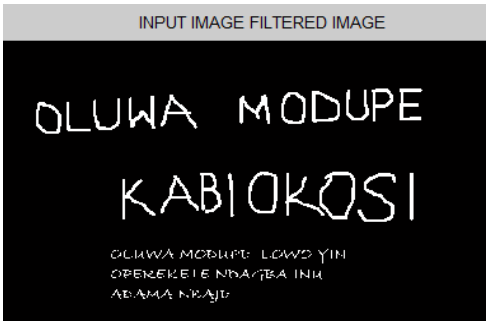

Figure 2(c). Input image text after filtered

Figure 3(a). OCR handwritten Yoruba character knowledge base

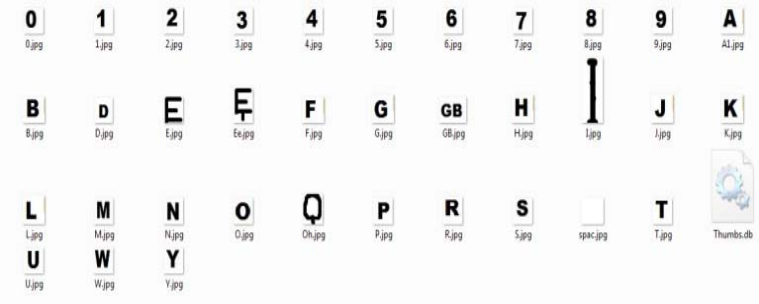

Figure 3(b). OCR typed Yoruba character knowledge base

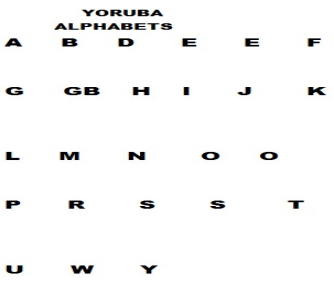

Figure 4(a). OCR input of a scanned image text document

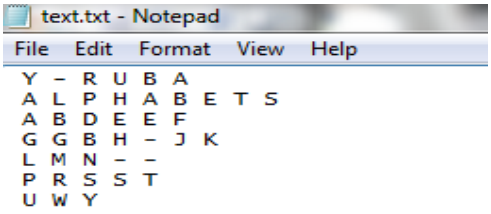

Figure 4(b). OCR output of the scanned image text document in 4a

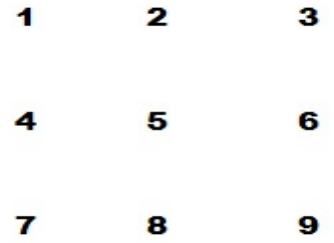

Figure 5(a). OCR input of a scanned image text document

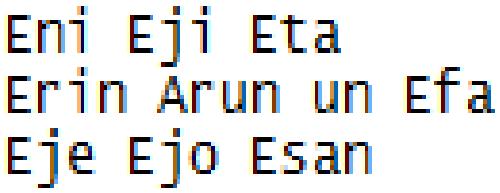

Figure 5(b). OCR output of the scanned image text document in 5a

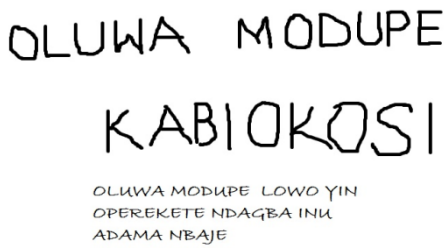

Figure 6(a). OCR input of a scanned handwritten image text document

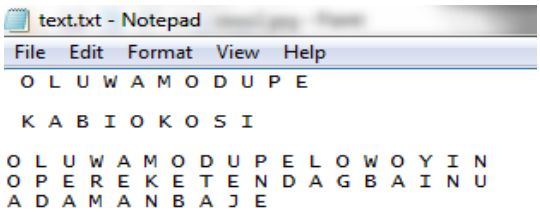

Figure 6(b). OCR output of a scanned handwritten image text document

4. Test and Discussion

The OCR system was subjected to different set of input text images in order to determine its recognition efficiency. The test was carried out on both typed and handwritten input texts. The input images as shown in Figure 4(a), 5(a) and 6(a) are different set of input texts created using the paint brush as pen and paint text which represent handwritten and typed Yoruba texts respectively. The outputs of the OCR system for the input text image are shown in Figure 4(b), 5(b) and 6(b).The test results were quite impressive. It was observed from the OCR output in Figure 4(b) that characters *I and O* were the only characters not recognized. This shows an accuracy of 86% for the typed text with execution time of 112 char/sec recognition rate. Also, for input text in Figure 5(b) it was observed from the OCR output in Figure 5(b) that all the English numbers were correctly recognized and converted to the Yoruba numerals. This showed accuracy of 100% for the numerals recognition and conversion. The OCR system output in Figure 6(b) which represents OCR output for the handwritten input text in Figure 6(a), also recorded an accuracy of 100%. It was observed that the developed Yoruba OCR system's performance unit is independent and constant for handwritten and typed text images of different sizes. Also, the result showed that the developed OCR system more effectively recognized numerals than alphabets.

References

[1] Isaac, Adejoju O. The Search for a Yoruba Orthography since the 1840s Obstacles to the Choice of the Arabic. *Sudanic Africa*. 2003; 77-102.
[2] A Jain, Karu K. Page Segmentation Using Texture Analysis, Pattern Recognition. 2006; 743-770.
[3] Kundu A, MITRE Corp, McLean, Hines T, Phillips J, Huyck BD. Arabic Handwriting Recognition Using Variable Duration HMM. *ICDAR. Ninth International Conference on*. Document Analysis and Recognition. 2007.
[4] Dueire Lins, R, Pereira Silva, G, Gomes e Silva AR. Assessing and Improving the Quality of Document Images Acquired with Portable Digital Cameras. *ICDAR. Ninth International Conference on.* Document Analysis and Recognition. 2007; 2: 569-573.
[5] Yin, Xu-Cheng, Sun, Jun, Naoi, S, Fujimoto K. A Multi-Stage Strategy to Perspective Rectification for Mobile Phone Camera-Based Document Images. 2007.
[6] Sun, Jun, Huang, Kaizhu, Hotta Y, Fujimoto K. Degraded Character Recognition by Complementary Classifiers Combination. *ICDAR. Ninth International Conference on.* Document Analysis and Recognition. 2007.
[7] Pratap RL, Satyaprasad L, Sastry A. Middle Zone Component Extraction and Recognition of Telugu. *ICDAR Ninth International Conference on Document Image Document Analysis and Recognition*. *2007,*
[8] Mohammed, Ali Q. Template Matching Method for Recognition Musnad Characters based on Correlation Analysis. *ACIT.* 2011.
[9] Huang, Kaizhu, Sun, Jun, Hotta, Y, Fujimoto K. An SVM-Based High-accurate Recognition Approach for Handwritten Numerals by Using Difference Features. *ICDAR, Ninth International Conference on Document Analysis and Recognition*. 2007; 589-593.
[10] RMK Sinha, et.al. HybridContextual Text Recognition with String matching. *Pattern Analysis and Machine Intelligence (PAMI).* 1997; 915-925.
[11] Fukunaga K. *Introduction to Statistical Pattern Recognition*. 1990.
[12] Huang, Gary, Learned-Miller, Erik, McCallum, Andrew. Crytogram Decoding for Optical Character Recognition.
[13] Kamaljit, Kaur, Balpreet, Kaur. Character Recognition of High Security Number Plates Using Morphological Operator. *International Journal of Computer Science & Engineering Technology. IJCSET.* 2013; 4(5).
[14] Lin, Shang-Hung. An Introduction to Face Recognition Technology. *Informing Science special issue on Multimedia Informing Technologies*. 2000; 3(1).
[15] Nadeem, Danish, Rizvi, Saleha. *Character Recognition Using Template Matching*.
[16] Nawaz, Tabassam, Hassan, Syed Ammar, Naqvi, Shah, Rehman, Habibur, Faiz, Anoshia. Optical Character Recognition System for Urdu (Naskh Font) Using Pattern Matching Technique.
[17] Qing, Chen, Emi, Petriul M. Optical Character Recognition for Model-based Object Recognition Applications.
[18] Saqib, Rasheed, Asad, Naeem, Omer, Ishaq. Automated Number Plate Recognition Using Hough Lines and Template Matching. *Proceedings of World Congress Engineering and Computer Science WCECS.* 2012.
[19] Ullmann JR. *Application of Pattern Recognition*. CRC Press, Inc., 1987.