Hidden Markov Model Approach for offline Yorùbá Handwritten Word Recognition

Jumoke F. Ajao [*1] **Stephen O. Olabiyisi**[2], **Elijah O. Omidiora**[2] **and Oladayo O. Okediran**[2]

[1] *Department of Computer Science, Kwara State University, Malete,*

[2] *Department of Computer Science & Engineering, Ladoke Akintola University of Technology, Ogbomosho*

Type of Article

Abstract

This paper presents a recognition system for *Yorùbá* handwritten words using Hidden Markov Model(HMM).The work is divided into four stages, namely data acquisition, preprocessing, feature extraction and classification. Data were collected from adult indigenous writers and the scanned images were subjected to some level of preprocessing, such as: greyscale, binarization, noise removal and normalization accordingly. Features were extracted from each of the normalized words, where a set of new features for handwritten *Yorùbá* words is obtained, based on discrete cosine transform approach and zigzag scanning was applied to extract the character shape, underdot and the diacritic sign from spatial frequency of the word image. A ten(10) state left-to-right HMM was used to model the *Yorùbá* words. The initial probability of HMM was randomly generated based on the model created for *Yorùbá* alphabet. In the HMM modeling, one HMM per each class of the image feature was constructed. The Baum-Welch re-estimation algorithm was applied to train each of the HMM class based on the DCT feature vector for the handwritten word images. Viterbi algorithm was used to classify the handwritten word which, gave the corresponding state sequences that best describe the model. Our experiments reported the highest test accuracy of 92% and higher recognition rate of 95.6% which, indicated that the performance of the recognition system is very accurate.

Keywords: keywords:Handwriting Recognition, Yorùbá Corpus, Yorùbá Language, Database Schema, Database Interface

2016 Mathematics Subject Classification: 53C25; 83C05; 57N16

**Corresponding author: E-mail: ajaofalilat@yahoo.com*

1 Introduction

The problem of handwritten Yorùbá word recognition (HYWR) has not been widely researched for a long time in spite of its potential in many applications such as document analysis, mail sorting, document archiving, commercial form-reading, office automation and so on Cheriet, Kharma, Liu and Seun (2007). A few known methods have been proposed among which is Ibrahim and Odejobi (2011) who presented a system for Character Recognition using bayesian and decision tree classifiers on Yorùbá six upper case letters.

The possibility of adapting modern Information Technology (IT) to our specific and unique needs is motivating. Hence, in the process of developing the resources that are needed to make modern IT relevant to African languages, there is need to incorporate indigenous language into the recognition system. so as to save Yorùbá language from total extinction. Due to differences in the nature of scripts in different languages, separate works have been done for various languages (Cheriet, Kharma, Liu and Seun, 2007). Diacritics marks are rare in English but are common occurrences in Yorùbá language. Yorùbá character is tonal, which makes its recognition more difficult than that of English language sequence of characters. Using under dots make some Yorùbá letters 'special' such as in ẹ, ọ, ṣ. One Yorùbá letter is a digraph represented as 'gb'. Several Yorùbá letters include vowel diacritical. The presence or absence of vowel diacritical indicates different meanings (Bamgbose, 1976). For example: bá (marked with an acute Accent) refers to 'meet'; ba, which is unmarked, signifies 'weave' and bà, marked with a grave accent refers to 'impinge upon'. Diacritical marking are essential to differentiate between possible meanings.

The diacritical marking may be ignored in handwritten unless the words are isolated, and this introduces additional difficulty in our recognition task. As removal of any of these dots will lead to misinterpretation of the character, efficient pre-processing techniques have to be used in order to deal with these dots without removing them and changing the identity of the characters. (Falola and Genova, 2006). The diacritical marking may be ignored in handwritten unless the words are isolated, and this introduces additional difficulty in our recognition task. As removal of any of these dots will lead to a misinterpretation of the character, efficient pre-processing techniques have to be used in order to deal with these dots without removing them and changing the identity of the character.(Falola and Genova, 2006)

Handwriting recognition can be divided into on-line and off-line recognition according to the format of handwriting inputs: In off-line recognition, only the image of the handwriting is available, while in the on-line case temporal information such as pen tip coordinates, as a function of time, is also available (Plamondon and Srihari, 2000). Besides, the basic handwriting recognition algorithms, various post processing methods are proposed to improve the accuracy of recognition. The techniques for on-line and off-line recognition may be quite different, but the post processing methods are possibly similar for both of them (Plamondon and Srihari, 2000). Many applications require off-line HWR capabilities such as bank processing, mail sorting, document archiving, commercial form-reading, office automation, and so on.

This paper focuses on Yorùbá handwritten word recognition. And it uses discrete cosine transform to extract features of Yorùbá handwritten words. Our algorithm showed a better recognition rate. The remainder of this paper is structured as follows: Section 2 presents the structure of Yorùbá language while Sections 3 presents the review of related works, section 4 describe the proposed method in terms of pre-processing and feature extraction; Section 5 describes the HMM classification process in details; experimental results are presented in Section 6. The paper ends with conclusions and suggestions for further work.

Many well known techniques used to recognize handwriting have been reported in Cheriet, Kharma, Liu and Seun (2007). Experiments shows that the method of Hidden Markov Model(HMM) adopted for European handwriting recognizers are not appropriate for Yorùbá handwritten word. An unconstrained Yorúbà handwriting is naturally challenging for off-line recognition systems since the words include many under dots and diacritical marks which change the meaning of the word

(Makhoul, Schwartz, Lapre and Bazzi, 1998; Ibrahim and Odejobi, 2011). There is need to model the parameters of HMM that will be appropriate for Yorúbà handwritten words with correct orthography, correct tone mark and under dot in lower case letters and upper case letters.

2 *Yorùbá* Language

The Yorùbá language, is spoken natively by over thirty million people in West Africa, primarily in Nigeria and in the neighbouring countries of the Republic of Benin and Togo. In Nigeria, Yorùbá speakers reside in the Southwest region in states like Ọ́yọ̀, Ogun, Ọsun, Ondo, Ekiti, Lagos, Kogi and Kwara. Yorúbà language is also spoken in the Diaspora in places like Cuba, Brazil, and the Caribbeans (Falola and Genova, 2006; Ojo, 2007). Yorùbá is a tonal language with three tones: high, mid and low. The high tone is indicated by an acute accent (á, é, ẹ́, í, ó,and ú). The mid tone is not marked and the low tone is marked with a grave acute (à, è, ẹ̀, ì, ò, ẹ̀ and ù). Some characters have dots at the bottom to indicate different tonal pronunciation. The need for inclusion of Yorúbà language in IT, to save the Yorúbà language from going to extinction has motivated this research work.

Yorùbá is written in a similar way to English with few differences (Amin, 2000). Yorùbá alphabet differs from English in omitting C, Q, V, X, and Z but adding Gb (one letter), Ẹ, Ọ, and Ṣ. Comparing Yorùbá and English sequence of characters, there are some similarities between the two. Both English and Yorùbá are read and written from left to right, have capital and small letters, have spaces between the words and share some common upper case and lower case letters. Despite these similarities, there are some significant differences between the two. English alphabet comprises of twenty-six capital and twenty-six small letters, but, Yorùbá alphabet has twenty-five capital letters and twenty-five small letters. Sequences of one or more of these letters represent the basic spoken units of a word or a complete word. This sequence of letters is known as a grapheme and the sound unit it represents is a phoneme. Yorùbá alphabet can be broken down into three tiers:

i. Tonal Tier

ii. The Under dot Tier

iii. The Character Tier and The diagraph

The tonal tier can be categorized into three tiers. The high-tone, the mid -tone and the low-tone. The under dot tier are the characters that carries the under dot sign. The character are set of characters that are common to both European alphabet and the Yorùbà alphabet except some few differences: these information is represented in Table 1.

Table 1: The *Yorùbá* alphabet

Toner Tier					
High-Tone(Acute sign)	Mid-tone	Low- tone(grave tone)			
(á, é, ẹ́ , í, ó, ọ́ and ú)	i j	è,ẹ̀ ,ì, ò, ọ̀ and ù			
The Under dot Tier					
Ẹẹ	Ọọ	Ṣṣ			
The Character Tier					
Aa	Bb	Dd	Ee	Ff	Gg
Hh	Ii	Jj	Kk	Ll	
Mm	Nn	Oo	Pp	Rr	Ss
Tt	Uu	Ww	Yy		

3 Review of Related works

Ibrahim and Odejobi (2011) presented a system for character Recognition using Bayesian and decision tree classifiers on Yorùbá Six upper case letters. Handwriting words from different writers were collected. The handwritten word collected were passed to Bayesian classifier and the decision tree algorithm was used to determine diacritic sign and under dot. 94.44% recognition rate was achieved.

El-Yacoubi, Gilloux, Sabourin and Seun (1999) designed a system to recognize unconstrained handwritten words. The preprocessed word images were divided explicitly into a sequence of segments and two feature sets were extracted from the sequence of segments. The word models were made up of the combination of appropriate letter models and an HMM-based interpolation technique was used to combine the two feature sets. The recognition system developed considered two rejection mechanisms depending on whether or not the word image belong to the lexicon. Experiments were carried out on 4,313 French city name images real mail envelopes, and 93% recognition rate was achieved.

Arica and Yarman-Vural (2001) proposed an analytic scheme,for cursive handwriting recognition. The global parameters, such as slant angle, baselines, and stroke width and height were extracted.The Segmentation approach finds character segmentation paths by combining gray scale and binary information. And HMM was employed to label shape recognition and rank the character candidates. The estimation of feature space information and HMM ranks are combined in a graph optimization problem for word-level recognition. The performance of the system was tested using 2,000 words of the database of Lancester-Oslo/Bergen corpus of cursive handwriting. The scheme yielded a very high recognition rate.

Femwa (2012) presented an hybrid feature extraction techniques using Geometrical and Statistical features for recognizing online characters. A hybridized classification model was developed to train the neural network using modified counter propagation and modified Optical back propagation learning algorithms. The results obtained showed the learning rate parameters variation had a positive effect on the network performance. The research worked on single Latin character in isolation and 96% recognition rate was achieved.

Gunter and Bunke (2003) combined three classifiers with different architectures for handwritten word recognition. A new ensemble method with several base classifiers was applied and the results of the ensemble method were compared to the results of the combination of the three classifiers. An increase of 2.98% was obtained with the best combination scheme. The authors also show that by using classical ensemble methods, such as Bagging and AdaBoost, the performance could also be increased. However, the best performance was achieved with a new ensemble method proposed by the authors, which distinguishes it from classical ensemble methods by the fact that it uses several base classifiers, rather than just a single one.The performance of the new ensemble method was 1.5% higher than the best combination of the base classifiers, 2.94% higher than the classical ensemble methods, and 4.48% higher than the best base classifier.

Jain, Duin and Mao (2000) summarized and compared some of the well known methods like template matching, statistical, syntactic or structural, neural networks used in various stages of a pattern recognition system and identified research topics and application which were at the forefront of this challenging field.

Kazim, Babak and Ehsanollah (2014)presents an offline Farsi/Arabic handwritten recognition using a contourbased method that provides rich set of discriminate characteristics to improve the recognition rate. A fast contour alignment was used to extract the feature point an a reduced lexicon based on loci features was applied on the features of binary images. The recognition system was validated on IBN SINR database which yielded recognition rate of 91.08

Koerich, Leydier, Sabourin and Seun (2002) proposed two level HMM decoding algorithm to deal with large vocabulary handwriting. The authors propose a non-heuristic fast decoding algorithm which is based on hidden Markov model representation of characters. The decoding algorithm breaks up the computation of word into two levels: state level and character level. Given an observation

sequence, the two level decoding enables the reuse of character likelihoods to decode all words in the lexicon, avoiding repeated computation of state sequences. The proposed scheme yielded a better recognition results.

Ajao, Jimoh, and Olabiyisi (2012) presented a handwritten recognition system of destination address on envelopes using Artificial Neural Network. The Back propagation algorithm was used to classify the word using a segmentation-based approach. This was deployed on different destination address on envelopes, considering the thirty-six states, 96% recognition rate was achieved which gave a promising result.

Impedovo, Ferrante and Modugno (2009) presented a new approach for Handwritten Word Recognition based on Hidden Markov Model theory and the sliding window technique. The new approach uses specific singularity markers to support the recognition phase: the Static Marker and the Dynamic Marker. Moreover, different strategies for sliding window step were considered: Regular Step and Progressive Step. Experimental results shows the improvements obtained for basic word lexicon recognition were reported in the paper.

Cheng-Liu (2009) presented a recognition system of Indian and Arabic handwritings. To test the promise of existing handwritten numeral recognition methods and provide new benchmarks for future research, their paper presented some results of handwritten Bangla and Farsi numeral recognition on binary and grayscale images. For recognition on grayscale images, it proposes a process with proper image pre-processing and feature extraction. The paper justified the benefit of recognition on grayscale images against binary images, compared some implementation choices of gradient direction feature extraction, some advanced normalization and classification methods.

Azizi, Farah and Sellami (2010) worked on Arabic Recognition system using Ensembles Classifiers, which focused on studying the effect of fusion methods on global system performance, he used the Diversity measure and individual accuracy classifier for selecting the best classifier, the best among the classifier was then used for the classification purpose. The experimental results presented were encouraging and open their perspective in the domain of classifiers selection.

Wen-Li and Mei-Huei (2015) the paper presents normalization process for handwriting recognition with scribbling data of different resolutions collected from different devices, such as touch screens and tablets. The normalization algorithms aimed at position, scale and rotation invariant in order to standardize non-uniform handwriting results from all sorts of users. The recognition process starts with identifying the bound of a handwriting. The cropped bound is centered to the origin and then scaled to a default size without producing undesirable distortions. Image skewing is handled by sampling data image of multi-angles through rotation transformation to produce extra learning artifacts. Down-sampling is employed by mingling neighborhood pixels into blocks to improve learning and recognition speed. The empirical studies show that this proposed standardization approach can yield a high degree of accuracy.

4 Hidden Markov Model for Recognition

Hidden Markov Model was chosen as the classifier due to the following: it provides methods for incremental learning of new classes and it performs automatic segmentation of string. This means that new handwritings can be added to the database without recomputing the representations of all other learned handwritings. HMM has also proved effective for a number of other tasks, such as handwriting recognition and sign language recognition. Because each HMM uses only positive data, they scale well; since new words can be added without affecting the already trained HMM patterns. Rabiner (1989).

The Hidden Markov Model (HMM) is a variant of a finite state machine having a set of hidden states, Q, an output alphabet (observations), O, transition probabilities, A, output (emission) probabilities, B, and initial state probabilities, π The current state is not observable. Instead, each state produces an output with a certain probability (B).

In the training phase,the Baum-Welch algorithm was model to optimize the training data through an iterative process. a variant of the Expectation Maximization (EM) algorithm, was used to maximize the observation sequence probability P(O | λ) of the chosen model $\lambda = (A, B, \pi)$. Where parameters A, B and π, denote matrix of transition probabilities, observation probabilities, and initial states probabilities. Rabiner (1989)

In the testing phase, a modified Viterbi algorithm was used to search for the corresponding state sequence that optimize model probability that matches the word given the input feature vector. In order to use HMM for recognition, an observation sequence is obtained from the test signal and then the likelihood of each HMM generating this signal is computed using the forward-backward algorithm. The HMM which has the highest likelihood then identifies the test signal using Baum-Welch algorithm and the state sequence which maximizes the probability of an observation, this is done using the Viterbi algorithm, which is a simple dynamic programming optimization procedure. Rabiner (1989)

Given a model of(Yorùbá word | λ) and an observation P(O | λ), how do we get an observable sequence that will maximize the model? and how do we determine the best observable sequence that will satisfy the model. Rabiner (1989) Five things that are required to solve this problem:

1. *N: The number of states to represent the model*
2. *M: The number of distinct observable state*
3. *Aij: The transition probability where*

$$a_{ij} = P[q_t + 1 = S_j | q_t = S_i, 1 \leq i, j \leq N] \tag{4.1}$$

4. $B_i(k)$ *the observable probability*

$$b_i(k) = P\{V_k \;\; at \;\; t | q_t = S_j\} \tag{4.2}$$

5. *the initial probability* π

$$\pi_i = P\{q_i = S_i\}, 1 \leq i \leq N \tag{4.3}$$

The problems required to be solved using HMM are:

1. *The Evaluation problem*
2. *The Re-estimation problem*
3. *The Decoding problem*

4.1 The Evaluation Problem

In order to apply HMM for Yorùbá word Recognition, HMM efficiently compute the model given an observation of Yorùbá word. The forward-backward algorithm was deployed to solve the evaluation problem Rabiner (1989):

4.1.1 The Forward algorithm:

Initialization:

$$\alpha_1(i) = \pi_i b_i(o_1) \tag{4.4}$$

Induction

$$\alpha_t(j) = [\sum_{i=1}^{N} \alpha_t - 1(i)] b_j(O_t + 1), 1 \leq t \leq T - 1, \;\; 1 \leq J \leq N \tag{4.5}$$

Termination

$$P(X|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{4.6}$$

where $\alpha_i(t)$*is the partial observation sequence of the HMM Model*

4.2 The backward algorithm

The backward algorithm is stated below: Initialization

$$\beta_T(i) = 1 \tag{4.7}$$

Induction

$$\beta_t(i) = \sum_{i=1}^{N} a_{ij} b_j(O_t + 1)\beta_{t+1}(j) \quad t = T - 11 \leq i \leq N \tag{4.8}$$

where $\beta_t i$ is the partial observation symbol of state q_t at time t-1,... 1.

4.3 The Training Process

The Baum-Welch algorithm was used to re-estimate the learning parameters to produce the expected model parameters that gives details representation of the word. The expected model parameters (A, B and π) generated with the extracted features, was used to train the Hidden Markov Model and the HMM model generated a transition matrix that gives the probability of the observation matrix. The Baum-Welch algorithm (Baum, Petrie, Soules, and Weiss, 1970), was used to train the Hidden Markov where re-estimation of the model parameter is done in order to get the optimal corresponding state sequences that actually gives the model. Rabiner (1989) The Baum-Welch algorithm is as follows:

$$L = \prod_{y=1}^{Y} P(O^y|\lambda) \tag{4.9}$$

where 0^y = the yth observation sequence, $\overline{a_{ij}}$ is the expected transition matrix and $\overline{b_j(k)}$ is the expected observation sequence.

$$\overline{a_{ij}} = \frac{\sum_y^Y \frac{1}{P(O^y|\lambda)} \sum_{t-1}^T \sum_{i=1}^N \alpha_{t-1}^y(i) a_{ij} b_j(o_t) \beta_t^y(j)}{\sum_y^Y \frac{1}{P(O^y|\lambda)} \sum_{t-1}^T \sum_{i=1}^N \alpha_t^y(i) \beta_t^y(i)} \tag{4.10}$$

$$\overline{b_j(k)} = \frac{\sum_y^Y \frac{1}{P(O^y|\lambda)} \sum_{t-1}^T \sum_{i=1}^N \delta(O_t^y = v_k) \alpha_{t-1}^y(i) a_{ij} b_j(o_t) \beta_t^y(j)}{\sum_y^Y \frac{1}{P(O^y|\lambda)} \sum_{t-1}^T \sum_{i=1}^N \alpha_{t-1}^y(i) a_{ij} b_j(o_t) \beta_t^y(j)} \tag{4.11}$$

4.4 The Classification Problem

The algorithm used to carry out the classification problem is the use dynamic programming techniques that was used to get the corresponding state sequences that optimizes Observation sequence given the model.

4.4.1 The Viterbi algorithm

Viterbi algorithm is a dynamic programming algorithm that chooses the best state sequence which maximizes the likelihood of the state sequence for the feature vector of the observation sequence. Let$\delta_t i$ be the maximal probability of state sequence of the length t that end in state i and produce the t first observations for the given model.

$$\delta_t(i) = maxP(q(1), q(2), q(3); o(1), o(2), o(t)|q(t) = q_i \tag{4.12}$$

Initialization:

$$\delta_1(i) = P_i b_i(o(1)) \tag{4.13}$$

$$\Psi_i(i) = 0, i = 1, ...N \tag{4.14}$$

This is a formal extension of the below recursion to:
Recursion:

$$\delta_t(j) = max_i[\delta_{t-1}(i)a_{ij}]b_j(o(t)) \tag{4.15}$$

$$\Psi_r(j) = argmax_i[\delta_{t-1}(i)a_{ij}] \tag{4.16}$$

Termination:

$$P^* = max_i[\delta T(i)] \tag{4.17}$$

$$P^* = q_T^* = argmax_i[\delta_T(i)] \tag{4.18}$$

5 The Recognition Process

Considering the fact that the Yorùbá medical pathology word image contains fields that belong to different categories such as the (Toner tier, Under dot tier, Character tier, and Digraph tier), it is difficult to process the entire Yorùbá word image at the same time efficiently. The main procedures in the proposed recognition system consist of pre-processing stage to remove extraneous things that may affect the performance of the recognition system. The preprocessed handwritten words were subjected to feature extraction techniques to extract relevant pattern from the word, such as the under dot and the diacritic signs. The feature extraction techniques was done holistically to extract the loop, the ascender, the descender of the Yorùbá word. The extracted features were subjected to the classification algorithm. Detailed of the processes is explained in this section.

6 Data Acquisition

Fifty(50) medical pathology words were obtained from the English medical dictionary which were translated to its Yorùbá equivalent. The translated Yorùbá handwritten medical words were written by 200 literate indigenous writers, which amounts to total of ten thousand(10,000) words of medical pathology words. Each of these ten thousand(10,000) words were replicated using different resolutions, formats and sizes. The handwritten image was scanned using 300 dpi resolution, 200 dpi resolution and 100 dpi resolution using different format such as JPEG, bitmap, png and tif at different sizes. The aim was to determine the robustness of recognition system developed on various image of different format so as to determine its effects on the recognition rate, the computational cost and the convergence ratio. The scanned handwritten words were used to create the Yorùbá handwriting database.

7 *Yorùbá* handwritten word Recognition

The following steps were deployed to recognize the offline Yorùbá handwritten words.

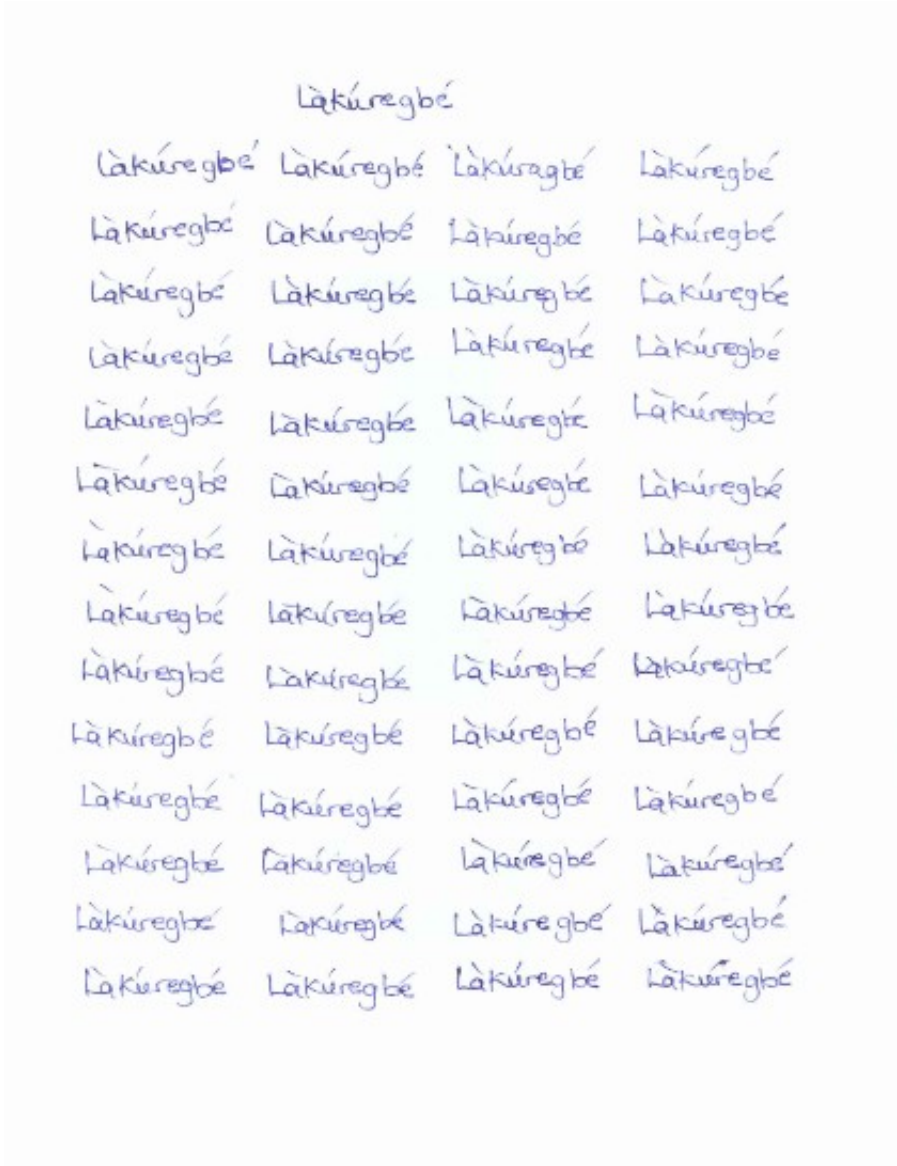

Figure 1: Sample handwriting of *Yorùbá* word *Làkúregbé*

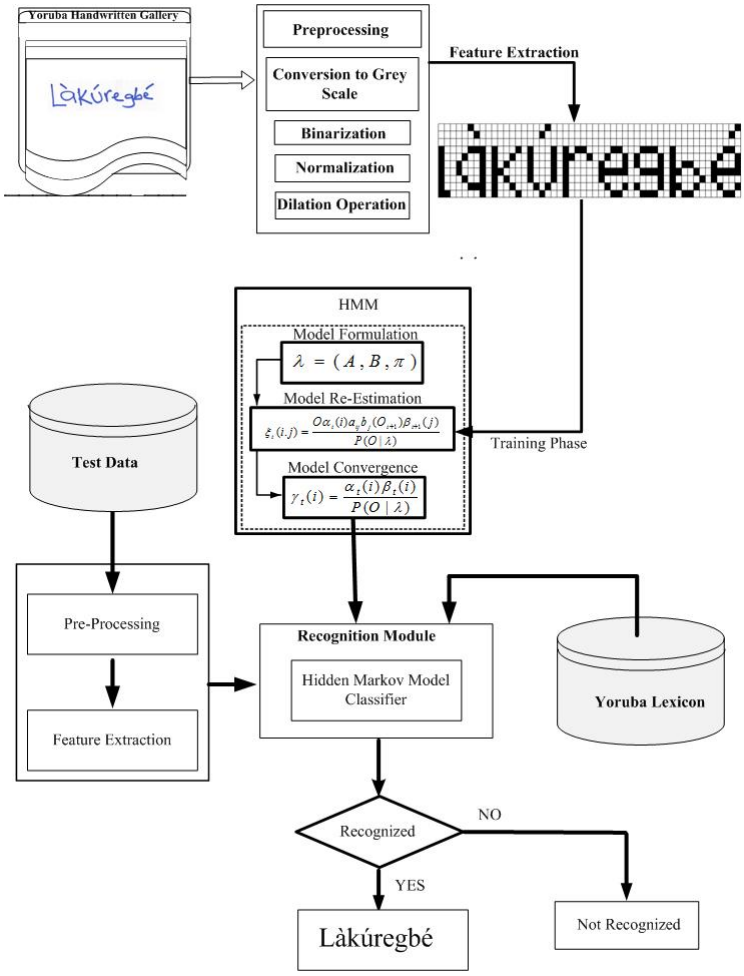

Figure 2: The framework architecture of the Recognition System

7.1 Preprocessing

The acquired data were subjected to a number of steps to make it usable in the descriptive stages of character analysis, the handwritten acquired were scanned and the scanned images were subjected to some pre-processing stages in order to extract appropriate invariant features that was used by the classifier for the recognition system. The image pre-processing stages are as follows:

Stage 1: *Conversion of the scanned image to grayscale Initially, the samples handwritten images were in RGB format, this stage converts the true color image RGB to greyscale intensity image by eliminating the hue and saturation information in the sample handwriting. The greyscale images were converted to its binary form and the samples handwriting were normalized by dividing the dimensions of each image by 255.*

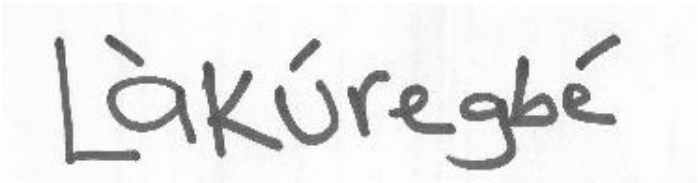

Figure 3: The Grayscale image of Làkúregbé handwritten word

Figure 4: The Binary image of *Làkúregbé* handwritten word

Figure 5: The Binary Gradient image of *Làkúregbé* handwritten word

7.2 Feature Extraction

Discrete cosine transform is an orthogonal transform method proposed by Ahmad, Natarajan and Rao (1974). DCT has been widely applied in image processing research since it was proposed. Discrete cosine transform can be derived from discrete Fourier transform (DFT) If a function is both real and even, then its Fourier series contains only cosine terms. Various approaches to the DCT is a well-known signal analysis tool used in compression due to its compact representation power (Chadha , Vaidya and Roja, 2011). Two-dimensional DCT can be defined as follows:

$$X(u,v) = \frac{2}{N}\sum_{n_2=0}^{N-1}\sum_{n_1=0}^{N-1} c(u)c(v)x(n_1,n_2)cos\left[\frac{(2n_1+1)u\pi}{2N}\right]cos\left[\frac{(2n_2+1)v\pi}{2N}\right] \tag{7.1}$$

where $n_1,\ n_2,\ u,\ v\ = 0, 1, \cdots,\ N-1$.
$x(n_1,\ n_2)$ *represent coefficients data in original block after applying DCT.*

$$c(u) = c(v) = \begin{cases} \frac{1}{\sqrt{2}} & \textit{if}\ u = 0,\ v = 0 \\ 1 & \textit{otherwise}\ . \end{cases}$$

The inverse discrete cosine transform (IDCT) can be defined as follows:

$$x(n_1,n_2) = \frac{2}{N}\sum_{v=0}^{N-1}\sum_{u=0}^{N-1} c(u)c(v)X(u,v)cos\left[\frac{(2n_1+1)u\pi}{2N}\right]cos\left[\frac{(2n_2+1)v\pi}{2N}\right] \tag{7.2}$$

DCT helps separate the image into parts (or spectral sub-bands) of differing importance (with respect to the image's visual quality). The Discrete Cosine Transform was used as feature extraction techniques that transform the handwritten image from its spatial domain to its frequency domain. it carries out the analysis of the frequency image in a zigzag scanning method to extract the peculiar features of the handwriting. A holistic feature extraction was carried out that extracts the ascender, the descender, hook , loop and diacritic sign on the word was carried out. The ascender and the descender was used to extract the shape of characters, the acute accent, the grave accent and the under dot that make up Yorùbá handwritten words were extracted. No explicit segmentation of characters was carried out. The default representation of the DCT algorithm uses 7×7 *array of integers. Yorùbá handwritten words comprises of diacritic mark on some of the vowel characters, which makes the array of integers to be extended to* 16×16 *array of integers to capture the diacritic marking and the underdot on some characters.*

7.3 Modelling *Yorùbá* Letters with Hidden Markov Model

Yorùbá handritten word was represented by a 10 - states left-to-right Hidden Markov Model as illustrated in Figure 7. Ten(10) was chosen based on the formulation of the Yorùbá alphabet. The model was formulated based on the assumptions that, handwritten can contain ordinary character, character with grave accent, character with acute accent, character with under dot, character with under dot and grave accent, character with under dot with acute accent and additional empty space unintentionally added in the course of writing and additional three states were added that captured lower case letters, upper case letters and title case letters to fully represent Yorùbá alphabet which amount to the ten left-to-right HMM . Fourty(40) observation sequence was used to represent the Yorùbá alphabet, which was used to train the Hidden Markov Model. Each of the state has loop, which denotes that a state can go into itself depending on the structure of the letter. Figure 6 shows the HMM topology representation of the Yorùbá handwritten word làkúregbé.

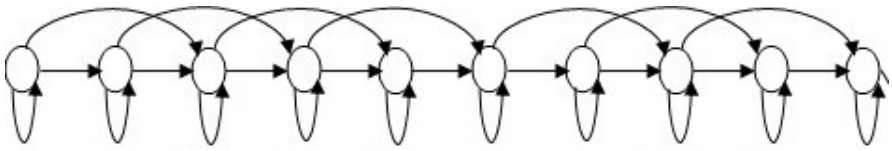

Figure 6: A 10-state Left-To-right Hidden Markov Model used to model *Yorùbá* alphabet

7.4 The Hidden Markov Model Initialization for *Yorùbá* word:

All these initial values were formulated:
The initial $probability \quad \pi_i = P\{q_i = S_i\}$
Transition matrix = $[10 \quad \times \quad 10]$ *matrix*
Observation sequence =[40], there are 25 characters that make up Yorùbá alphabet the additional fourteen plus one observations is gotten from seven instances of the vowel characters with acute sign and seven instances of vowel characters with grave accent and the under dot, character with under dot and grave accent, character with under dot, grave accent and an empty space.
The size of the observation sequence is [10 x 40] observations for each states of the Yorùbá alphabet model.
B_i *is a back pointer keeping track of the partial path leading to state i at time t. Once it reaches state N, it recursively apply the back pointer B to retrieve the entire optimal sequence* S^*

8 *Yorùbá* Recognition Process

A left-right with skipping options HMM topology was used to model the Yorùbá alphabets.The multiple structure captured both characters in uppercase and lowercase letters. The recognition process uses an holistic recognition process where the handwritten Yorùbá words was captured holistically without considering the segmentation of each of the characters that make up the Yorùbá words.The descender, ascenders, the holes and hooks of handwritten Yorùbá words is captured which gave detailed representation of the word to model Yorùbá handwritten word in its real word scenario. A flat static lexicon of fifty words were generated for the lookup table created used in the recognition module.

The skipping option was chosen to model Yorùbá Handwritten word recognition system. Since, a word can contain ordinary character or it could be of any of the seven instances of vowel characters with grave accents or acute accents and an empty space unintentionally added during the course of writing. Each of the states captured the different scenario that make up Yorùbá handwritten words.

The recognition module reliably recognized the handwritten medical pathology words or misclassified Yorùbá handwritten word with ambiguity. Figure 7 illustrates the basic modules in the handwritten medical pathology word recognition system. The Lookup table (ID representation): A lookup table was created, that is a set of handwritten words (U) and a set of typewritten words (V) such that each element x of a set U is associated to a single element y of a set V. In this case, every element of V represents the image of a single element of U as illustrated in Figure 7. The Recognition mode:

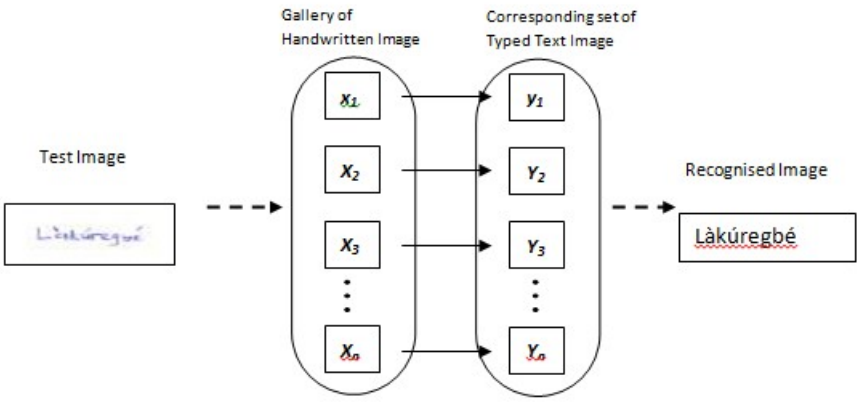

Figure 7: The block diagram of the recognition process

8.1 Algorithm to Recognize *Yorùbá* handwritten Word using Hidden Markov Model

1. *Define a set of Word database or class for modeling the Yorùbá words*
2. *Generate a random values for π, A and B from the model of Yorùbá letters.*
3. *Initialize $q_t = s_t$*
4. *For each class, collect a training data set gotten from the extracted features to train the HMM*
5. *Compute the observation sequence to produce the given model.*
6. *Based on each training, solve the estimation problem to optimize the HMM parameters.*
7. *Evaluate Probability of the observation sequence given the model to find the most probable path that will produce the observation given the unknown features)*

9 Experimental Result

The recognition of Yorùbá handwritten words was performed on the Yorùbá handwritten database created by the Authors. Ten thousand(10,000) medical pathology words were obtained from literate indigenous writers. The samples handwritten were scanned using an OCR device in order to digitized it. The digitized handwriting images were subjected to some level of preprocessing to enhance the quality of the Yorùbá handwritten images. The digitized images were normalized and each of the dimensions of the images were divided by 255. Discrete Cosine Transform was used to extract relevant information that represents the handwritten words. Hidden Markov Model was used to classify the handwritten words. The handwritten words were stored in the database and the lexicon for each of the word was created to serve as the lookup table for the handwritten words. The handwritten images stored in the database were used to train the HMM classifier. In order to validate the YHWRS developed, handwritten word images were selected from the database created, and it was tested on the recognition module, that captured the preprocessing, feature extraction and the classification process. The input data were subjected to preprocessing, features extraction of the preprocessed image were carried out and compared with the trained features vector. If the two feature sets are

similar, it look up for appropriate word in the look-up table and display the correspondence digital equivalent of the handwritten word if found, or brings a False Negative result.

$$Precision = \frac{FN}{TN + FN} \tag{9.1}$$

where FN is the False Negative, TN is the True Negative.

- *The recognition rate is calculated as:*

$$percentage\ of\ recognition\ = \frac{Number\ of\ recognized\ word}{Total\ Number\ of\ Test\ data} \times 100 \tag{9.2}$$

Table 2 shows the recognition accuracy with each of the handwritten words. Some percentage of the handwritten samples were used to train the HMM and the remaining handwritten words were also used to the test the YHWRS , in oder to test its reliability. Table 3 shows the Recognition accuracy for open-test validation and Table 4 shows the recognition accuracy for close-test validation of the Yoruba handwritten words.

Table 2: Recognition Accuracy of the OCR *Yorùbá* words for lowercase letters with open-test letter words

Handwritten Image	Train data	Test data	Recg	NotRecg
ikọ́	2	30	28	2
òtútú	2	30	28	2
ẹ̀fọ́rí	2	30	30	0
agara	2	30	30	0
sòbìà	2	30	27	3
ibà	2	30	29	1
àjàkálẹ̀	2	30	28	2
àrùn	2	30	29	1
àtọ̀sí	2	30	29	1
ìlẹ́gboná	2	30	28	2
sísunú	2	30	28	2
alákásá	2	30	28	2
segede	2	30	28	2
làkúrègbé	2	30	30	0
iwárápá	2	30	30	0

(Makhoul, Schwartz, Lapre and Bazzi, 1998; Ibrahim and Odejobi, 2011).

Table 3: Recognition of the *Yorùbá* handwritten medical pathology words of Title case

Handwritten Image	Train data	Test data	Recog	NotRecog
Òtútù	2	30	28	2
Efori	2	30	30	0
Agara	2	30	30	0
Sobia	2	30	27	3
Iba	2	30	29	1
Ajakale	2	30	28	2
Arun	2	30	29	1
Atosi	2	30	29	1
Iko	2	30	28	2
Sisunu	2	30	28	2
Segede	2	30	28	2
Ete	2	30	28	2
Ewo	2	30	29	1
Irindo	2	30	27	3
warapa	2	30	30	0
Jedijedi	2	30	28	2
Lakuregbe	2	30	30	0
Lapalapa	2	30	30	0
Agan	2	30	29	1
Alakasa	2	30	29	1
Arunsu	2	30	28	2
Atogbe	2	30	29	1
Eebi	2	30	30	0
Egbo	2	30	30	0
Ofinkin	2	30	28	2
Ogbeenu	2	30	27	3

Table 4: Recognition of the *Yorùbá* medical pathology words with open-test validation

Handwritten Image	Train data	Test data	Recog	NotRecog
Otutu	2	10	7	3
Efori	2	10	8	0
Agara	2	10	10	0
Sobia	2	10	8	2
Iba	2	10	9	1
Ajakale	2	10	9	1
Arun	2	10	10	0
Atosi	2	10	9	1
Iko	2	10	7	3
Sisunu	2	10	9	1
Segede	2	10	9	1
Ete	2	10	10	0
Ewo	2	10	9	1
Irindo	2	10	6	4
warapa	2	10	9	1
Jedijedi	2	10	9	1
Jedojedo	2	10	8	2
Lakuregbe	2	10	7	3
Lapalapa	2	10	9	1
Agan	2	10	9	1
Alakasa	2	10	8	2
Arunsu	2	10	8	2
Atogbe	2	10	8	2
Eebi	2	10	9	1
Egbo	2	10	8	2
Ofinkin	2	10	7	3
Ogbeenu	2	10	6	4

Table 5: Recognition Accuracy of the OCR *Yorùbá* words using close-test validation

Handwritten Image	Train data	Test data	Recg	NotRecg
Otutu	2	10	10	0
Efori	2	10	9	1
Agara	2	10	10	0
Sobia	2	10	10	0
Iba	2	10	10	0
Ajakale	2	10	9	1
Arun	2	10	10	0
Atosi	2	10	10	1
Iko	2	10	10	0
Sisunu	2	10	10	0
Segede	2	10	10	0
Ete	2	10	10	0
Ewo	2	10	10	0
Irindo	2	10	9	1
warapa	2	10	10	0
Jedijedi	2	10	10	0
Jedojedo	2	10	10	0
Lakuregbe	2	10	10	0
Lapalapa	2	10	10	0
Agan	2	10	10	0
Alakasa	2	10	10	0
Arunsu	2	10	10	0
Atogbe	2	10	9	1
Eebi	2	10	10	0
Egbo	2	10	10	0
Ofinkin	2	10	10	2
Ogbeenu	2	10	10	0

9.1 Performance of *Yorùbá* HWR systen with other handwritten Word Recognition System

Table 6: Performance Evaluation of *Yorùbá* medical pathology words with other Handwritten word Recognition using HMM and other classifier extracted from Document Analysis Journal

Method	Classifier	Lexicon Size	RR	Test Set	Database
(Bunke, 1995)	HMM	150	98.4	3,000	Words (ENG)
(Mohamed, 1996)	HMM-DP	100	89.3	317	City names (Eng)
(Knerr,1998)	HMM-NN	30	92.9	40,000	LA word (FR)
(Guillevic, 1998)	HMM-K-NN	30	86.7	4,500	LA Word (ENG)
(EL Yacoulbi, 1999)	HMM	100	96.3	4,313	City names (FR)
(EL Yacoulbi, 1999)	HMM	1,000	88.9	4,313	City names (FR)
(Kim, 2000)	HMM-MLP	32	92.2	2,482	LA Word (ENG)
(Freitas 2001)	HMM	38	77.0	2,387	LA Word (POR)
(Oliveier Jr, 2002)	MLP	12	87.2	1,200	Month Word
(Xu,2002)	HMM-MLP	29	85.3	2,063	Month Word (FR/ENG)
(Kundu,2002)	HMM	100	88.2	3,000	Postal Word (FR/ENG)
(Our proposed scheme)	HMM	150	95.6	1,540	*Yorùbá* HW database

10 Conclusion

In this paper, we applied HMM classifier to recognize Yorùbá handwritten medical pathology words. It was observed that, HMM classifier with DCT algorithm works effectively for the recognition of Yorùbá handwritten words. Tables 2, 3, 4 and 6 show the recognition accuracy with each of the handwritten words. A few number of handwritten words generated from the Yorùbá handwritten database were used to train the HMM and the remaining handwritten words were used to test the reliability of the YHWRS. Table 2 shows the Recognition accuracy for lower case letters, Table 3 shows the recognition accuracy for Title case handwritten words. It was also observed that the YWRS developed shows insignificant results in the recognition of the handwritten words using both open-test validation and close-test validation as shown in Table 4 and Table 6. Table 6 shows the performance evaluation of Yorùbá handwritten word recognition system(YHWRS) developed with other Handwritten word recognition system using HMM and other classifiers. The YHWRS developed was compared with other HWRS, and it was observed that the YHWRS achieved a recognition rate of 95.6% using 150 lexicons, test data of 1540 handwritten word images which was validated on Yorùbá handwritten database created.

References

Ahmad, M., Natarajan , T. and Rao, K. R. (1974). Discrete Cosine Transform, IEEE Trans. Computers, 1974, pp. 90-94.

Chadha , A. R., Vaidya, P. P., Roja, M. M. (2011). Face Recognition Using Discrete Cosine Transform for Global and Local Features, Proceedings of the International Conference on Recent Advancements in Electrical, Electronics and Control Engineering (IConRAEeCE) IEEE Xplore: CFP1153R-ART.

Adeyanju, I., Fenwa, O., and Omidiora, E. O. (2014). Effect of Non-Image Features on Recognition of Handwritten Alpha-Numeric Characters. International Journal of Computers and Technology, 13.

Ajao, J. F., Jimoh, R. G., and Olabiyisi, S. O. (2012). Handwritten Address Dsetination Recognition using Neural Networks. Journal of Science, Technology, Mathematics and Education, 9(1):70–91.

Al-amri, S., Kalyankar, N., and Khamitkar, S. (2010). Image segmentation by using edge detection. International Journal of Computer Science and Engineering, 03(2):804–807.

Alkhateeb, J., Ren, J., and Jiang, J. . (2011). Performance of Hidder Markov Model and Dynamic Bayesian network classifiers on handwritten Arabic word Recognition. Pattern Recognition, pages 680–688.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3):175185.

Amin, A. (2000). Recognition of printed arabic text based on global features and decision tree learning techniques. Pattern Recognition, 33(8):1309–1323.

Arica, N. and Yarman-Vural, F. (2001). An overview of character recognition focused on off-line handwriting. IEEE Trans. Systems Man Cybernet, 31(2):216-233.118.

Assabie, Y. and Bigun, J. (2003). Lexicon-based offline recogntion of Amharic words in unconstrained handwritten text . In Pattern Recognition, 19th International Conference, pages 8-11, Tampa, Florida, USA. IEEE.

Awal, A., Mouchere, H., and C., V.-G. (2010). A hybrid classifier for handwritten mathematical expression recognition. In Document Recognition, page 10. IEEE.

Azizi, N. and Farah, N. (2012). New dynsmic ensemble of classifiers selection approach based on confusion matrix for arabic handwritten recognition. 2012 International Conference on Multimedia Computing and Systems IEEE, 14(2):308-313.

Azizi, N., Farah, N., and Sellami, M. (2010). Off-line Handwritten Word Recognition System using Ensemble of Classifier Selection and Fusion. Journal of Theoretical and Applied Information Technology, 14(2):141–150.

Bahoura, M. and Simad, Y. (2012). Serial Combination of Multiple classifiers for automatic whale calls recognition. Pattern Recognition, pages 9986–9993.

Bamgbose, A. (1976). Yorùbá Orthography, Ibadan University Press, pp. 15-27.

Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer.

Breiman, L. (1996). Bagging Predictors. Machine Learning, 24(2):123-140.

Bunke, H. (2003). Recognition of Cursive Roman handwriting-Past, Present and Future. Document Analysis and Recognition, (39):9986-9993.

Cheng-Liu, L. (2009). A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters. Journal of Pattern Recognition, 42(12):3287-3295.

Cheriet, M., Kharma, N., Liu, C., and Seun, C. (2007). Character Recognition Systemss: A Guide for Students and Practitioners. John Wiley. 119

Desai, A. (2010). Gujarati handwritten numeral optical character reorganization through neural network. Elsevier, pages 2582-2589.

Dinh, C., Duin, R., and Loog, M. (2012). A study on semi-supervised dissimilarity representation. Pattern Recognition (ICPR), 9(2):2861–2864.

Duda, R., Hart, P., and Stark, D. (2000). Pattern Classification. John Willey, Canada, 2nd edition.

Duda, R., Hart, P., and Stork, D. (2001). Pattern Classification. John Wiley, second edition.

El-Yacoubi, A., Gilloux, M., Sabourin, R., and Seun, C. (1999). An HMM-based approach for off-line unconstrained handwritten word modelling and recognition. IEEE, TPAMI, 21(8):752–760.

Falola, T. and Genova, A. (2006). The Yorùbá Transition: Values and Modernity. Journal of Asian and African Studies, 45(6):576–577.

Femwa, O. (2012). Development of a writer-independent online handwritten Character Recogntion using modified hybrid neural network. PhD thesis, Ladoke Akintola University, Ogbomoso, Department of Computer Science and Engineering, Faculty of Engineering and Technology.

Fenwa, O., Omidiora, E., and Fakolujo, O. (2012). Development of a feature Extraction Technique for Online Character Recognition System. Innovative Systems and Design Engineering, 3:10–23.

Gunter, S. and Bunke, H. (2003). Ensembles of classifiers for handwritten word recognition. International Journal of Document Analysis and Recognition, 5(2):224–232.

Heutte, L., Paquet, T., Moreau, J., Lecourtier, Y., and Olivier, C. (2000). A structural/ statistical feature based vector for handwritten character recognition. In Pattern recognition letters , volume 19, pages 629–641, North-Holland.

Ibrahim, A. and Odejobi, O. (2011). A System for the Recognition of handwritten Yorùbá Characters. In AGIS 2011, Obafemi Awolowo University, Ile-Ife, Nigeria. AGIS.

Impedovo, S., Ferrante, A., and Modugno, R. (2009). HMM Based Handwritten Word Recognition System by Using Singularities. 12th International Conference on Document Analysis and Recognition, pages 783–787.

Jain, R., Duin, R., and Mao, J. (2000). Statistical Pattern Recognition-A review. Pattern Recognition, 22(1):4–27.

Kahan, S., T., P., and Baird, H. (2009). On the Recognition of Printed Characters of Any Font and Size. Pattern Analysis and Machine Intelligence, IEEE transaction, 9(2):274–288.

Koerich, A., Leydier, Y., Sabourin, R., and Seun, C. (2002). A hybrid large vocabulary handwritten word recognition system using neural networks with hidden Markov Models. Internal Journal of Document Analysis and Recognition, pages 99–104.

Makhoul, J., Schwartz, R., Lapre, C., and Bazzi, I. (1998). A Script-Independent Methodology for Optical Character Recognition. Pattern Recognition, 31(9):1285– 1294.

[34]Ojo, O. (2007). The Yorùbá Transition: Values and Modernity. Journal of Asian and African Studies, 54(2):151–152. 121

Omidiora, E., Adeyanju, I., and Femwa, O. (2013). Comparison of Machine Learning Classifiers for Recognition of Online and Offline Handwritten Digits. Computer Engineering and Intelligent Systems, 4(13):2222–2863.

Plamondon, R. and Srihari, S. (2000). On-line and Off-line handwriting recognition. IEEE transactions on pattern analysis and Machine Intelligence, 22(1):63–84.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. volume 77 of Proceedings of the IEEE, pages 257–285.

Ranawana, R. and Palade, V. (2006). Multi-classifier Systems:Review and Roadmap for Developers. International Journal of Hybrid Intelligent Systems, 3(1):35–61.

Saritha, B. and Hemanth, S. (2009). An Efficient Hidden Markov Model for Offline Handwritten Numeral Recognition. InterJRI Computer Science and Networking, 1:7–12.

Senthikumaran, N. and Rajesh, R. (2009). Edge Detection Techniques for Image Segmentation. International Journal Recent Trend in Engineering, 2(2).

Shanon, C. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 273:379–423.

Theodoridis, S. and Koutroumbas, K. (2009). Pattern Recognition. Academic Press, San Diego.

Verma, R. and Ali, J. (2012). A-Survey of Feature Extraction and Classification Techniques in OCR Systems. International Journal of Computer Applications and InformationT echnology, 1(3):2278–7720. 122

Waleed, H. and Nikola, K. (1999). The concept of hidden markov model in speech recognition. PhD thesis, Knowledge Engineering Lab.Information Science, Department, University of Otago , New Zealand.

Wozniak, M., Grana, M., and Corchedo, E. (2014). A survey of multiple classifier systems as hybrid systems. Information Fusion, Science Direct, 16:3–7.

Yi-Lu, M. and Hong, G. (2000). Automatic Feature Selection an hybrid statistical approach. Pattern Recognition, 34.

Kazim, F., Babak N. A., and Ehsanollah, K.(2014) . A fast and accurate contour-based method for writer-dependent offline handwritten Farsi/Arabic subwords recognition. IJDAR, pages:17:181203 .

Wen-Li, W., and Mei-Huei, T. (2009). A Normalization Process to Standardize Handwriting Data Collected from Multiple Resources for Recognition. Elsevier, pages :61:402 409.

Zhou, X., Wang, D., and Liu, C. (2009). A robust approach to text line grouping in online handwritten Japanese documents. Pattern Recognition, pages 2077–2088. 123.