International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 9 – No.1, June 2015 – www.ijais.org

Yoruba Handwriting Word Recognition Quality Evaluation of Preprocessing Attributes using Information Theory Approach

Jumoke F. Ajao
Kwara State University
Malete, Ilorin

Stephen O.Olabiyisi
Ladoke Akintola University of Technology, Ogbomoso

Elijah O.Omidiora
Ladoke Akintola University of Technology, Ogbomoso

Odetunji O. Odejobi
Obafemi Awolowo University
Ile-Ife, Osogbo

ABSTRACT
This paper presents an approach to evaluate the quality of handwritten words using the set of features in the preprocessing stages. This is to determine the effect of various stages of preprocessing of the recognition of Yoruba handwritten characters. We demonstrate our methods using handwritten words samples of the domain of Yoruba medical terminology collected from indigenous Yoruba literate writers. Samples were digitized at 300dpi to facilitate much detail representation of the image dataset. We study the impact of entropy measure for an intuitive interpretation of the analysis of the handwritten words. From the experiment carried out, it was observed that the entropy measure of handwritten word is higher than the typewritten word. This implies that the Information content of the handwritten word is affected by perturbations which need to be removed using appropriate image preprocessing tools to obtain low entropy measure which implies same information content as the original.

General Terms
Pattern Recognition, Handwritten word recognition, Image preprocessing and Information Theory.

Keywords
Yoruba, Entropy, Handwritten word, and Optical Character Reader.

1. INTRODUCTION
HANDWRITING recognition (HWR) emanates from the need for automated machine recognition of human written text or the ability of the computer to receive and interpret human handwriting [1], [2], [3]. It has been a popular area of research for a few decades under the purview of pattern recognition and image processing [4] – [7]. A fundamental challenge in automated handwriting recognition is that recognition problems that appear to be simple for most people may in fact be quite difficult when transferred to machine domain [8], [9].Despite over five decades of intensive research, handwriting recognition continues to be an active area of research because of many unsolved fundamental theoretical problems as well as a rapidly increasing number of applications that can benefit from it [8], [9].

Handwriting recognition can be divided into on-line and off-line recognition, according to the format of handwriting inputs [10], [11]: In offline 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. Many applications require off-line HWR capabilities such as bank processing, mail sorting, document archiving, commercial form-reading, office automation, and so on.

Accent marks or diacritics are rare in English but are common occurrence in Yoruba [12]. Yoruba characters are tonal, which makes its recognition more difficult than that of printed English sequence of character. The writing system from right-to-left predominantly uses only consonants in their written forms. Vowels can be added, usually by means of diacritics. 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 (e.g. à), and the mid tone usually left unmarked. In some circumstances the mid tone is indicated with a 'macron'. These marks are usually placed on the vowels. Using dots makes some Yoruba letters special such as in 'ẹ, ọ, ṣ'. One Yoruba letter is a diagraph represented as '*gb'*.Several Yoruba letters include vowel diacritical. The presence or absence of vowel diacritical indicates different meanings [13]. Diacritical marking are essential to differentiate between possible meanings. However, the diacritical marking may be ignored in handwritten unless the words are isolated, and this introduces additional difficulty in a 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. The Yoruba Orthography is represented in the Table I:

Table I. The Yoruba Orthography

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
Underdot Tier	Ẹ ẹ Ọọ Ṣ ṣ	
Diagraph Tier	GB gb	
Character Tier	A B D E F G H I J K M N O P R S T U W Y a b d e f g h k l m n o p r s t u w y	

To strengthen the performance of handwriting recognition has been a major goal in HWR. Preprocessing is a vital step in handwriting recognition. However, the outcomes of incorporating all preprocessing stages are not always positive. Some favours preprocessing elements such as noise removal, slant or slope correction, binarization, etc. On the other hand, others argue that applying all the preprocessing stages might cause loss of vital information by removing or altering the originality [14].

This paper presents an approach to evaluate the quality of handwritten words based on the set of features that are used in the preprocessing stages. The use of these attributes allows very intuitive interpretation of the results and as a consequence provides solid foundation for implementation of Yoruba handwritten recognition system.We adapt information theory and related techniques in the development of a robust and accurate Yoruba word recognition system. The entropy scheme is then used to analyze the handwritten words where each observation corresponds to the image contents of the dataset.

The organization of the paper is as follows. Section I contains Introduction, Section II contains the related work to handwritten recognition methods are briefly discussed. In Section III, Preprocessing steps are detailed. The proposed entropy analysis method is introduced in Section IV. Section V presents the subjective and objective experimental results. Finally, the conclusion is provided in Section VI.

2. RELATED WORKS

Researchers in handwriting recognition have been treating recognition in different languages as a different problem. Each language has its own character sets and the language specific syntax rules governing how those characters are to be preprocessed. In this section, we briefly review handwritten recognition related methods.

Ibrahim and Odejobi [15] developed a system to recognize handwritten character of Yoruba Upper case letters. The work presented an approach of Bayesian and decision tree. It separated character into three regions: the top, the middle and the bottom which denotes the toner, the main part of the character and under dot. The paper presented a recognition rate of 94.44%. The research work focused on six Yoruba upper case characters only. The research establishes the knowledge and challenges of Yoruba recognition which forms the background information to this work.

Günter and Bunke[16] combined three classifiers: HMM, ANN and K-NN with different architectures for handwritten word recognition. They use several base classifiers, rather than just a single one, to derive an ensemble. A new ensemble method working 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 using Bagging and AdaBoost, However, the best performance was achieved with a new ensemble method proposed by the authors, which was distinguished from classical ensemble. An increase of 2.98% was obtained with the best combination scheme. However, the best performance was achieved with a new ensemble method proposed by the authors, which is distinguished from classical ensemble methods by the fact that 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.

Olivera[17] presented a summary of the recent advances in terms of character, word, numeral string, and sentence recognition was presented. Also, the main new trends in the field of handwriting recognition were discussed and some important contributions were presented. The paper reviewed several work on character, word, numeral strings and sentence recognition. It laid foundation for researchers on the path of text recognition.

Impedovo[18] presented a new approach for Handwritten Latin 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. Experimental results shows the improvements obtained for basic word lexicon recognition.

Desai [19] dealt with an optical character recognition (OCR) system for handwritten Gujarati numbers. In their work a neural network was proposed for Gujarati handwritten digits identification. A multi layered feed forward neural network was suggested for classification of digits. The features of Gujarati digits were abstracted by four different profiles of digits. Thinning and skew- correction were done for preprocessing of handwritten numerals before their classification. The maximum success rate achieved is for the digit four and that is 96.23%.

Kessentini *et al.*[20] presented a multi-stream approach for off-line handwritten word recognition. The proposed approach combines low level feature streams namely, density based features extracted from two different sliding windows with different widths, and contour based features extracted from upper and lower contours. The multi-stream paradigm provides an interesting framework for the integration of multiple sources of information and is compared to the standard combination strategies namely fusion of representations and fusion of decisions. Significant experiments have been carried out on two publicly available word databases: IFN/ENIT benchmark database (Arabic script) and IRONOFF database (Latin script).The proposed approach achieved a recognition rate of 89.8% using a lexicon of 196 words. The research work focused on offline Latin handwritten recognition and 89.6% recognition was achieved.

Femwa[21] presented a hybrid feature extraction techniques using Geometrical and Statistical features Handwritten Character Recognition. A hybridized classification model was developed to train the neural network using modified counter propagation and modified back propagation learning algorithms. The geometric and statistical features were used to extract the global and local properties of characters and topological features with tolerance to variation style and distortion. The results obtained showed the learning rate parameter variation had a positive effect on the network performance. Ninety six percent (96%) recognition rate was achieved.

El Yacoubi [22] presented a system to recognize unconstrained handwritten words for large vocabularies. The handwritten recognition system developed considered two rejection mechanisms depending on whether or not the word image is belong to the lexicon. After pre-processing, a word image was divided explicitly into a sequence of segments and then two feature sets were extracted from the sequence of segments. Their word models were made up of the concatenation of appropriate letter models and an HMM-

based interpolation technique was used to optimally combine the two feature sets. Their experiments were carried out on 4,313 French city name images manually localized on real mail envelopes, ninety-three (93%) of recognition rate was achieved.

3. PREPROCESSING STEPS

The generic image pre-processing shown in Fig 1 involves the following series of stages that are carried out in order to extract important information from an image and thereby reducing unwanted information from the image so as to enhance the recognition process [5], [14].

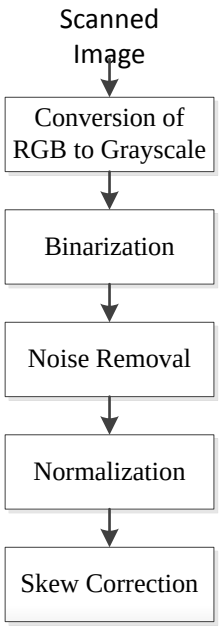


Fig 1. A Generic Image Preprocessing Tools

1. Conversion of RGB to Grayscale coverts image scanned into the computer to Grayscale image to enhance detail information needed from the image. Given that RGB value of a color is (R, G, B) such that R, G, and B are integers between 0 and 255. The grayscale weighted average, X, is given by the formula:

$$X = 0.299R + 0.587G + 0.114B \qquad (1)$$

2. Binarization is the transformation of a grayscale image into a black and white image through thresholding inspired by Niblack's algorithm [23]. It calculates a pixel-wise threshold by sliding a rectangular window over the gray level image given by the equation 1:

$$T_{Niblack} = m + k * s \qquad (2)$$

$$T_{Niblack} = m + k\sqrt{\frac{1}{NP}\sum (p_i - m)^2} \qquad (3)$$

$$= \mathrm{m} + \mathrm{k}\sqrt{\frac{\sum \mathrm{p_i^2}}{NP} - m^2} = m + k\sqrt{B} \qquad (4)$$

Where NP is the number of pixels in the gray image, m is the average value of the pixels pi, and k is fixed to -0.2.

3. Noise Removal eliminates noises/degradation that may be found in digital images. Scanned document images often have degradations like uneven contrast, interfering strokes, background spots, humidity absorbed by paper in different areas, and uneven backgrounds. Such degradations can be removed by several techniques such as [24]:

$$\frac{1}{9} \times \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \qquad (5)$$

$$\frac{1}{16} \times \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \qquad (6)$$

4. Normalization is the process of converting an image function $I_1(x,y)$ into the function $I_2(x,y)$ such that it retains all the relevant information of the original image. A class of normalization processes is described by [25] with the following relations between I_1 and I_2 as in equation (5).

$$\mathrm{I}_2(x_2, y_2) = GI_2(x_1, y_1) + B \qquad (7)$$

$$\begin{pmatrix} \mathrm{x}_1 \\ \mathrm{y}_1 \end{pmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \qquad (8)$$

5. Skew correction techniques can be categorized into direct, indirect and contour-oriented [25]. Equation (3) shows the direct method. Given the skew angle Θ, the direct method de-skews the image by rotating the black pixels by (-Θ) [26]:

$$\begin{pmatrix} \mathrm{x}' \\ \mathrm{y}' \end{pmatrix} = \begin{pmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{pmatrix} \bullet \begin{pmatrix} x \\ y \end{pmatrix} \qquad (9)$$

4. METHODOLOGY

Data were acquired from adult indigenous writers. The entropy of the Yoruba words was computed. The approach is aimed at assessing the intrinsic measure of some of the preprocessing stages. This is to provide information required in the preprocessing stage as a preparatory level for the handwritten Yoruba recognition system to be developed.

Samples of Yoruba pathology terms in handwritten and type written is shown in Fig 3.

Fig 3. (Left) handwritten (Right) type written word

The Data intrinsic value is measured using the Entropy equation (10):

$$H(x) = \sum_{w_i} p(w_i) \log_2 p(w_i) \qquad (10)$$

Where $p(w_i)$ is the proportion of x belonging to class w_i

At each level of the preprocessing stage, the data intrinsic measure is applied to ascertain the possibility of obtaining information that are ordered or closed the measure of the original information. The entropy specifies the level of uncertainty inherent in the preprocessed information (See Fig 4). The preprocessed features include Grayscale, Binarization and Binary gradient of the acquired handwritten dataset.

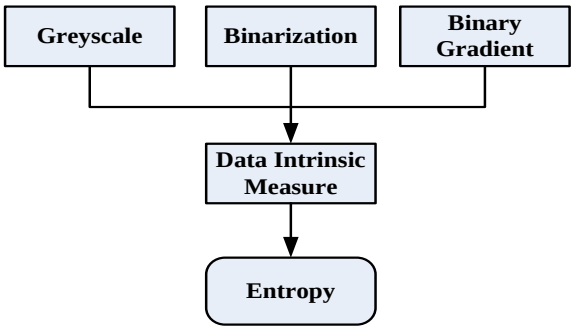


Fig 4. Data Intrinsic Measure on Preprocessed Dataset

5. RESULTS AND DISCUSSIONS

Table II, III and IV show the entropy and the count level of the handwritten samples after performing gray level preprocessing, binarisation and Binary gradient level.

On the basis of the experiment carried out using ImageJ software, differences in terms of entropy measure and the histogram distribution was noticed .This implies that the entropy of the typewritten word is observed to be lower than that of handwritten word. Which, corroborate the assertion of the Information theory that "the lower the entropy the richer the information contents [27]. It therefore implies that the information contents of the typewritten words are richer than their corresponding handwritten words. Consequently, it may be necessary to apply image pre-processing techniques to the handwritten word in order to achieve an entropy measure closer to the entropy measure of the original information, and this will enhance the likelihood of improving the accuracy of the handwritten Yoruba recognition system to be developed. After establishing the fact that, the entropy measure for the typewritten words is reduced compared to that of handwriting word, here are some of the pre-processing stages for the samples handwritten in order to reduce the entropy measure of the handwritten word.

The following pre-processing techniques were applied on the samples handwritten. The histogram spreads of the data were also taken. This was done to see the effect of pre-processing on the measure of Entropy of the handwritten word. Some the pre-processing techniques used are: Conversion to grayscale, Binarisation, Binary Gradient Mask and Binary Gradient Dilation.

In Table 2, 3 and 4, it was noticed that the value of entropy decreases for binary image, the count increases, and the value of count denotes that the information content increases as the entropy decreases. This implies that entropy is inversely proportional to information content. From experiment, it was observed that, the entropy measure of the handwritten decreases at each level of pre-processing which is tending towards the entropy measure of typewritten word. it was also observed that not all pre-processing techniques are required to be performed on handwritten image.

The histogram distributions and entropy of the handwritten words were taken, in order to see the effect of the pre-processing stage of the handwritten word.

From Fig 1 and Fig 2, it was also observed that, at the level of dilation, the entropy measure increases and the count level that denotes gain information content increase, which indicate that such level of pre-processing is not required since preprocessing of the handwritten samples is in successor, that is, the output of one preprocessed level will serve as input to the next level of preprocessing. Our major objectives lay emphasis on the entropy measure.

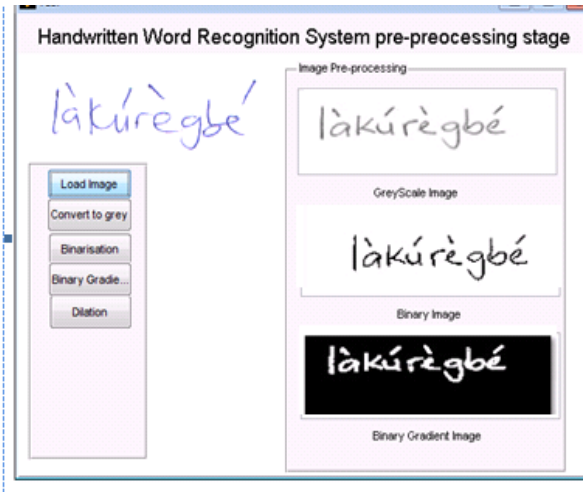


Fig 5. The Preprocessing Interface

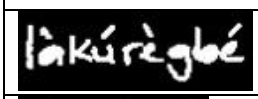
Table 2. The Entropy and histogram distribution of the handwritten Image in Grayscale

Yoruba Words	Count	Mean	Mode	Entropy
àkúrègbé	7176	247.886	4499	2.516
òtútú	4452	240.117	3296	2.31
àrùnsù	4343	247.984	3088	2.308

Table 3: The Entropy and histogram distribution of the handwritten Image in Binary format.

Yoruba Words	Count	Mean	Mode	Entropy
àkúrègbé	11088	243.665	8980	1.649
òtútú	5350	242.517	4085	2.095
àrùnsù	7467	244.844	6299	1.571

Table 4. The Entropy and histogram distribution of the handwritten in Dilation format

Yoruba Words	Count	Mean	Mode	Entropy
	14364	45.925	13706	2.287
	9842	27.649	7833	1.756
	10934	26.745	8474	1.872

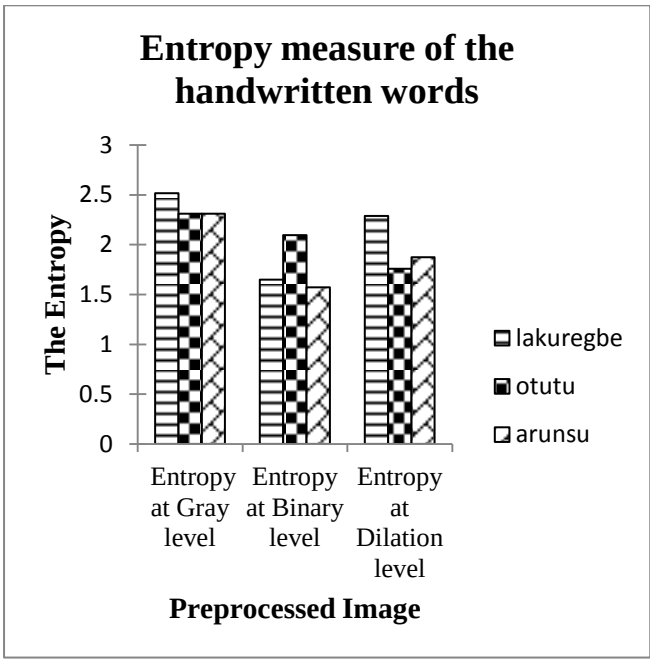


Fig 6. The entropy measure of the handwritten image at every level of preprocessing

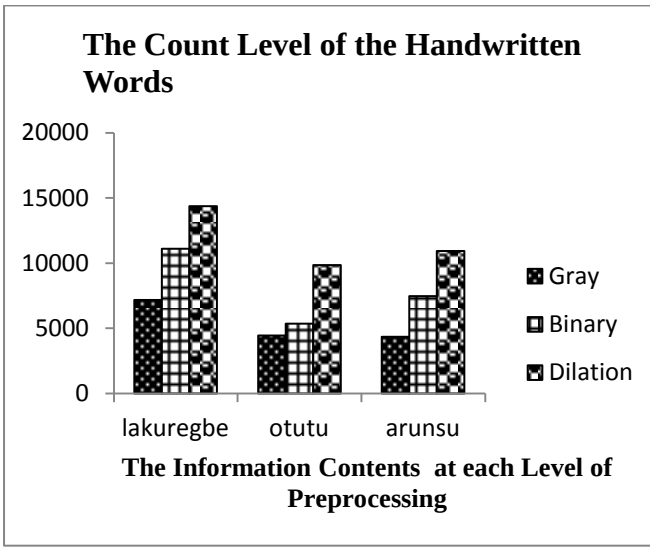


Figure 7. The count level of the handwritten image at every level of preprocessing

6. CONCLUSIONS

The objective of image preprocessing in handwriting recognition is to ensure original image restoration. In the present work, a novel method is devised to adapt the information theory and related techniques in the development of a robust and accurate Yoruba word recognition system. From experiment, it was observed that, the entropy measure of the handwritten decreases at each level of pre-processing which is tending towards the entropy measure of typewritten word.

The result implies that preprocessing stage to be deployed should be able to achieve an entropy measure closer to the entropy measure of the original Information.

Considering all the aspects discussed in the previous sections, the next steps to provide better Yoruba handwriting recognition systems are obvious. The recognition module will be able to determine the preprocessing stages that are required to be performed on the samples handwriting. This will enhance the performance of the Yoruba recognition to be developed.

7. REFERENCES

[1] Mori, S., Suen, C. and Yamamoto, K., 1992 "Historical Review of OCR Research and Development". Proceedings of IEEE, Vol. 80 No. 7, pp. 1029–1058.

[2] Saritha, B. S. and Hemanth S., 2009 "An Efficient Hidden Markov Model for Offline Handwritten Numeral Recognition," *InterJRI Computer Science and Networking*, Vol. 1, pp. 7-12.

[3] Marinai, S., Gori, M. and Soda, G. 2005 "Artificial Neural Networks for Document Analysis and Recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, Vol. 27, No.1, pp. 23 – 35.

[4] Jain, A. K., Duin, R. P., and Jianchang M. "Statistical pattern recognition: a review," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, Vol. 22, No. 1, pp. 4 – 37, January 2000.

[5] Duda, R. O. Hart, P. E and Stork, D. G. 2001 "Pattern classification," 2nd Edition, Wiley.

[6] Bishop, C. M. 2006 "Pattern Recognition and Machine Learning," Springer.

[7] Theodoridis S. and Koutroumbas, K. 2009 "Pattern Recognition", 4th Edition, *Academic Press*, San Diego,.

[8] Bunke, H. 2003 "Recognition of Cursive Roman handwriting – Past, Present and Future,"*Document Analysis and Recognition Seventh International Conference*, Edinburgh, pp. 448–459.

[9] Cheriet, M., Kharma, N. Liu, C-L. and Suen, C. "Character Recognition Systems: A Guide for Students and Practitioners," *John Wiley*, New York, November 2007.

[10] Plamondon R. and Srihari, S. N. 2000 "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey, "*IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1.

[11] Arica N. and Yarman-Vural, F. T., 2001 "An Overview of Character Recognition Focused on Off-Line

Handwriting," *Systems, Man, and Cybernetics, Part C: Applications and Reviews*, IEEE Transactions, Vol. 31, No. 2.

[12] Bamgbose, A. 1976 "Yoruba Orthography," Ibadan University Press, pp. 15-27.

[13] Omnigot 2013 "The Yoruba alphabets and it pronunciation,". Accessed at http://www.omniglot.com/writing/yoruba.htm.

[14] Lee, H. and Verma, B. 2012 "Binary Segmentation Algorithm for English Cursive Handwriting Recognition," *Pattern Recognition,Elsevier*, Vol. 45, No. 4, pp. 1306–1317.

[15] El-Yacoubi, A. Gilloux, M., Sabourin R. and Suen, C. Y. 1999 "An HMM-Based Approach for Off-Line Unconstrained Handwritten Word Modeling and Recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, Vol. 21, No. 8, pp. 752 – 760.

[16] Gunter, S. and Bunke, H. 2003 "Ensembles of Classifiers for Handwritten Word Recognition," *Document Analysis and Recognition*, Vol. 5, No. 4, pp. 224-232.

[17] Oliveira, L. S., Sabourin, R., Bortolozzi, F. and Suen, C. Y. 2002 "Automatic Recognition of Handwritten Numerical Strings: A Recognition and Verification Strategy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 11.

[18] Impedovo D. and Pirlo, G. 2014 "Zoning Methods for Handwritten Character Recognition: A Survey," *Pattern Recognition, Handwriting Recognition and other PR Applications*, Vol. 47, No. 3, pp. 969–981, Elsevier.

[19] Desai, A. A. 2010 "Gujarati Handwritten Numeral Optical Character Reorganization through Neural Network," *Pattern Recognition*, Vol. 43, No. 7, pp. 2582–2589, Elsevier.

[20] Kessentini, Y., Paquet T. and Hamadou, A. B. 2010 "Off-Line Handwritten Word Recognition Using Multi-Stream Hidden Markov Models" *Pattern Recognition Letters*, Vol. 31, No. 1, pp. 60 – 70.

[21] Femwa, O. D. 2012 "Development of a Writer-Independent Online, Handwritten Character Recognition System Using Modified Hybrid Neural Network Model," PhD. Thesis, Ladoke Akintola University of Technology, Ogbomoso,.

[22] Ibraheem, O. and Odejobi, O. A. 2011 "A System for the Recognition of Handwritten Yorùbá Characters," AGIS 2011 Ethiopia, ObafemiAwolowo University, Ile-Ife, Nigeria,. Retrieved from http://www.slideshare.net/aflat/a-system-for-the-recognition-of-handwritten-yoruba-characters.

[23] Niblack, W. 1986 "An Introduction Letters to Digital Image Processing," Prentice Hall, Englewood Cliffs.

[24] Abu-Mostafa, Y. S. and Psaltis, D. 1985 "Image Normalization by Complete Moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.7, No.1.

[25] Kwag, H. K., Kim, S. S. H. Jeony, S. H and G. S. Lee, 2002 "Efficient Skew Estimation and Correction Algorithm for Document Images", *Image and vision Computing*, Vol. 20, pp. 25-35.

[26] B. Yu, B. and Jain, A. K. 1996 "A Robust and Fast Skew Detection Algorithm for Generic Documents", *Pattern Recognition*, Vol. 29, Issue 10, pp. 1599-1629, October.

[27] Shannon, C. E., 1948 "A Mathematical Theory of Communication." *The Bell System Technical Journal*, Vol. 27, pp. 623–656.