A WORD-LEVEL YORUBA HANDWRITTEN CHARACTER RECOGNITION SYSTEM USING MODIFIED VISION TRANSFORMER

Supported by tetfund TERTIARY EDUCATION TRUST FUND

Mutiu Bolarinwa Falade[1], Ibrahim Adepoju Adeyanju[2]
[1]Department of Computer Engineering, Faculty of Engineering, Federal University Wukari, Nigeria,
[2] Department of Computer Engineering, Faculty of Engineering, Federal University Oye-Ekiti, Nigeria,
*Corresponding Author Email: mutiufalade@gmail.com

Received: April 16, 2024 Accepted: July 28, 2024

Abstract: Information on paper documents is accessible to those who are around where they are stored. Such important documents are not available in softcopy for search engines. A lot of Yoruba historical documents were handwritten before now. Automated Character recognition is relatively common for English language but not for Yoruba language due to the diacritics which lead to tonal difference in Yoruba text. This research aims to design and implement a Word-level Yoruba Handwritten Character Recognition system using modified Vision Transformer for character recognition. The Yoruba handwritten images consist of handwritten Yoruba alphabets from a newly created database locally acquired. The newly created dataset contains 6434 images with 40 classes. 6414 images were used as training dataset while 20 word images were used as testing dataset. Each image was resized to 72x72 pixels and then converted to grayscale. Noise was removed from each grayscale image using Otsu algorithm and bounding boxes for word image segmentation. Vision transformer was used for recognition and 21,692,904 trainable parameters were obtained from initial empirical experiments. Average segmentation, character-level recognition accuracy and testing time of 79.7%, 72.1% and 0.9s respectively were obtained. The developed system gave a better performance when evaluated with recognition accuracy and testing time which outperformed SVM and CNN classifiers when compared with existing studies. The developed system accommodates more Yoruba vocabulary and successfully recognized common Yoruba words with acceptable testing time. The developed system would be useful in archiving and digitizing Yoruba historical handwritten documents and teaching and learning in Yoruba language.

Keywords: Character**,** Handwritten, Modified, Recognition, Vision Transformer, Word-level, Yoruba

Introduction

Storing information on paper has typically been very difficult and expensive as they are exposed to theft, termite attack and fire. The problem of having very important documents in hardcopy and archiving of historical documents especially when such documents are not available and accessible to the public in softcopy (digital form) for search engines has been a serious concern. Such important documents are only accessible to those around where it is stored. Handwritten data on paper and printed documents are only accessible and available to those around where it is stored and several other people across the world might benefit from such documents and information. Digitizing such information helps search engines like Yahoo and Google to be able to make them accessible to the public (Adeyanju *et al*., 2014). Character recognition or Optical Character Recognition (OCR) is the recognition of characters processed optically. OCR is a field in pattern recognition and computer vision (Awel and Abidi, 2019). This is a technology that allows conversion of documents from scanned paper documents or images captured by a digital camera into editable and searchable data (Geetha *et al.*, 2022). Character recognition is the conversion of textual character images (handwritten or printed text) into computer or machine readable formats that can be edited using word processing tools. Optical Character Recognition of handwritten documents is still a challenging research area because of the unstructured and variable writing style of different individuals (Pareek et al., 2020). A lot of Yoruba historical documents were handwritten before now. Yoruba handwritten text is time consuming in typing because of its diacritics. Hence, there is need for optical character recognition system for digitizing these documents.

Character recognition is relatively common for English language but still very little for Yoruba language. Yoruba language is one of the three major languages spoken in Nigeria. It is spoken by people in south-western Nigeria, Benin, Togo, United Kingdom (UK), Brazil and United States of America (Oyeniran and Oyebode, 2021). Diacritics and tonal in Yoruba text makes Yoruba character recognition difficult than that of English alphabet (Ajao *et al.*, 2018). Handwritten text recognition is a challenging task and more difficult than printed text due to the differences in writing styles of individuals. Many methods have been developed for this purpose but none of these methods assure to give 100% accuracy (Shetty and Heraje, 2017). Handwritten character recognition is grouped into online and offline handwritten character recognition depending on how data is acquired. Offline data are capture using pen and paper before being converted into electronic form as images while online data are typically written directly using modern devices like a digitizer and an electric pen (Ojumah *et al.*, 2018).

This research aims is to develop a word-level Yoruba handwritten character recognition system using modified vision transformer classifier that will convert Yoruba words into digital format (editable format). This research contributed to the body of knowledge through:

i. Development of a Yoruba handwritten character recognition system using modified Vision Transformer.
ii. Creation of a new dataset for Yoruba handwritten character recognition with diacritics.

Other aspects of this research are arranged as; Section 1 presents introduction, section 2 and 3 discussed related works, materials and methods, results and discussion were presented in section 4. Finally, section 5 gives conclusions.

Review of Related Works

Many scholars have put in so much effort in the development of handwritten character recognition for English, French, Chinese, and Arabic but few systems have been developed for Yoruba handwritten characters. Several techniques have been proposed for Yoruba Handwritten and printed text. Ding, Jin and Gao (2009) worked on handwritten recognition for Chinese words. The result shows that the recognition accuracy of the proposed holistic method significantly outperformed that of analytical approach with increase of recognition rate by 11.71%. Nasien *et al*. (2010) developed a handwritten character recognition using Support Vector Machine (SVM) classifier for recognition. The model achieved low recognition accuracy. Kumar et al. (2010) developed a handwritten digit

recognition system using mathematical morphology. The average recognition accuracy of all digits was above 90%. The average recognition accuracy of all digits was above 90%. The system was unable to recognized broken digits with large gap.
Omidiora et al. (2013), compared machine learning classifiers for recognition of handwritten digits. Four machine learning classifier were considered. Naives bayes, Instanced based learner, decision trees and Neural networks were compared for single digit recognition. The instanced based learner outperformed the other three classifiers with a accuracy of 97.28%. The comparison was limited to handwritten digits. The system was only able to recognize upper case English alphabets. Adeyanju et al. (2016), developed recognition system for typewritten characters using Hidden Markov Models (HMM). Three sets of type written dataset were used. The system showed a recognition accuracy of 97.24%, 94.88%, and 91.45% respectively for newly typewritten essay, old memo and old war letter. The recognition system recorded its best result at 0.8 thresholds. The system cannot recognize handwritten characters and formatted characters. Ajao et al. (2018) developed a Yoruba handwritten character recognition using freeman chain code and k-nearest neighbour classifier. The system recognition accuracy was 87.7% which outperformed other classifiers. Their system only focuses on twenty four upper case Yoruba alphabets. Oyeniran and Oyebode (2021) presented a technique for Yorùbá alphabets recognition system using deep learning. The model accuracy was 97.97%. The proposed system was able to only recognized Yoruba alphabets. It did not consider Yoruba words.

Oladele et al. (2020) developed an offline Yoruba handwritten word recognition system that uses geometric feature extraction techniques and a SVM classifier. Recognition accuracy ranges between 66.7% and 100%. The words used in the system vary from four to seven letters word. Twenty different words were used. The system can only recognize upper case Yoruba letters. Oyeniran and Oyebode (2021) developed a Yoruba handwritten character recognition system using AlexNet, a deep learning model. The developed system achieved recognition accuracy of 91.4% and average recognition time of 0.371372seconds. However, the system was unable to recognize the alphabets '**Ṣ**' and '**ṣ**' with recognition accuracy of 0%.

Sonara and Pandi (2021) proposed Handwritten Character Recognition using Convolutional Neural Networks. Preprocessing techniques used include noise removal using median filter, grayscale conversion and image thinning. Segmentation was done using character segmentation and Convolutional Neural Network was used as classifier. The system achieved 95% recognition accuracy using English Handwritten characters. The research is only limited to English characters. Kala (2022) proposed Handwritten Character Recognition using Neural Networks for Hindi, Telgu, Kannada, Malayam and English. Preprocessing processes used were noise removal, skew detection and correction and binarization. Three types of segmentation were used (lines, word and characters segmentation). Classification was done using Convolutional Neural Networks. However, one language can be recognized at a time. There is need to create data module and dataset for classification of any given language irrespective of its origin. Patil et al. (2022) proposed enhancing OCR on mixed text images with semantic segmentation. Image preprocessing, image segmentation, image text processing using OCR and digitizing of extracted information were employed. The developed system provided quality input with several image preprocessing stages to OCR which gives higher recognition accuracy. The model initialization was heavy in terms of computation. It was unable to handle noisy images (input) as the quality of the output dropped.

Transformers are a type of deep learning architecture based primarily upon the self-attention mechanism or modules that were originally used for sequence-to-sequence tasks. It was first used in the field of Natural Language Processing (NLP) on machine translation tasks (Han et al., 2022). Two major components namely; feed forward network and self-attention are found in transformer architecture. In this case, information flows in only one direction, which is from input layer to output layer. Existing approaches based on self-attention uses single-head or multi-head (transformer) designs for vision tasks (Salman et. al., 2022). Multi-headed self-attention is a variant of attention used in most transformers. Multi-headed self-attention uses multiple different self-attention modules in parallel. The encoder portion of transformer has many repeated layers of identical structure. The transformer encoder consists of alternating layers of multi-headed self-attention and feed forward neural network modules (Vaswani et al., 2017). Decoders are not relevant to vision transformers which are encoder-only-architecture. Similar to encoder, decoder has many layers with masked multi-head attention, multi-head encoder-decoder attention and feed forward Neural Network. Figure 1 shows the architecture of a transformer.

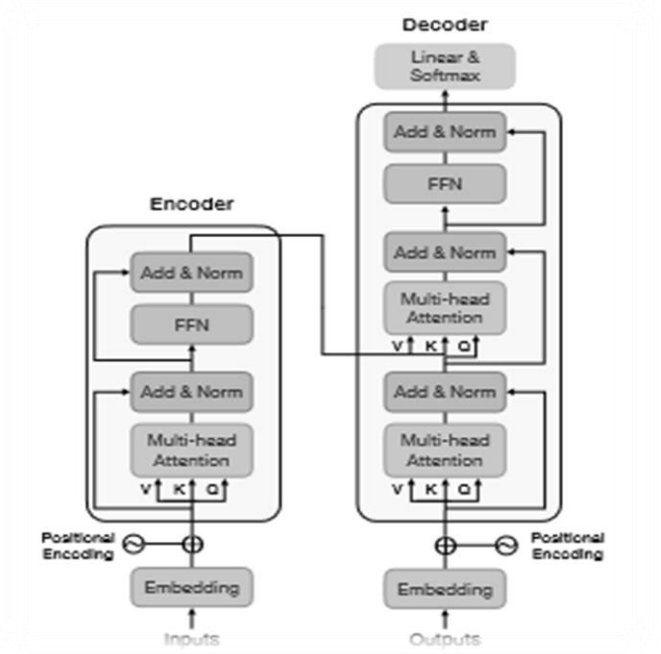

Figure 1: Architecture of Transformer (Bo et al.., 2022)

Deep learning research has achieved good results by using transformer architecture to computer vision tasks. Transformers that are applied in computer vision domain are referred to as Vision Transformer (ViT). ViT typically leverage on encoder-only transformer architecture. ViT contains three segments (Bo et al., 2022). They include patch and positional embedding, transformer encoder (feature extraction via stacked transformer encoder and Multi-layer Perceptron (MLP) head (classification head). Vision transformer models perform quite well relative to popular CNN variants on image classification task. ViT are now a viable and useful tool for deep learning experts. ViT is a model for image classification that uses a transformer-like architecture over patches an image where the image is split into fixed-size patches such as 16x16, 32x32, and 64x64 having patch size like 4, 6 and 8. Each of them is then linearly embedded, positional embedding is included and the resulting sequence of vectors is fed into an encoder (transformer layer). Figure 2 shows Vision Transformer architecture.

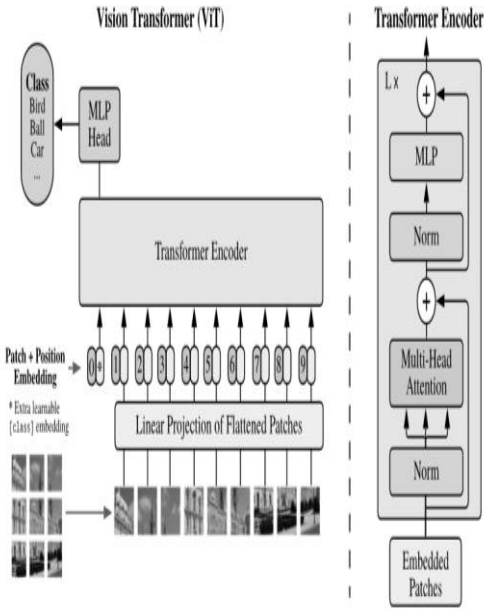

Figure 2: Vision Transformer Model overview (Dosovitskiy et al., 2021)

The research gaps and challenges for current Yoruba handwritten character recognition and other character recognition include limited character recognition, high computational cost and time and tonality of the language. This research focuses on the use of vision transformer, a deep machine learning algorithm for recognition of handwritten Yoruba text. Vision Transformer (ViT) is a deep learning model used for image classification that employs a transformer-like architecture over patches of images (Dosovitskiy et al., 2021). ViT has a higher accuracy when trained on large dataset and it uses multi-head self-attention in computer vision removing image-specific inductive biases. Literature review shows that when ViT is trained on sufficient data, it outperforms Convolutional Neural Networks (CNN) in the area of computational efficiency and accuracy.

This research study developed a model for word level Yoruba handwritten character recognition which would improve recognition accuracy, accommodate more vocabulary and testing time problems facing some of the existing developed systems.

Materials and Methods

Word-level Yoruba Handwritten Character Recognition was developed using python programming language on Google Colaboratory (Google Colab). A total of 6414 handwritten lower case character images were used as training data while 20 lower case word images were used as testing data. The stages are divided into five namely; data acquisition, image preprocessing, image segmentation, classification and system performance evaluation. Figure 3 shows the framework of the developed system.

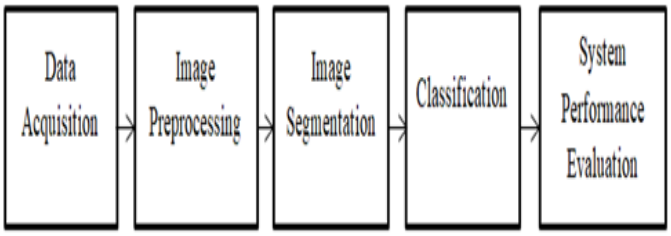

Figure 3: Framework of Word-level Yoruba Handwritten Character Recognition (WYHCR) System

Data acquisition

The first step in image processing techniques is image acquisition. Firstly, Images of Yoruba handwritten characters are publicly available in open-access github dataset which was downloaded from and available on https://github.com/oluwashina90/yoruba-handwritten–character-database. Secondly, another dataset was acquired locally from sixty volunteers. Forty writers wrote 40 classes of alphabets each and twenty writers wrote one word each. Thirdly, open-access dataset were combined with locally acquired dataset. The training dataset consists of 40 classes of characters. 34 classes of character which include a, à, á, b, d, e, é, è, ẹ, f, g, h, i, ì, í, j, k, l, m, n, o, ò, ó, ọ, p, r, s, ṣ, t, u, ù, ú, w and y contains 180 images each. 6 classes of character which include ẹ̀, ẹ́, ọ̀, ọ́, ǹ and ń consist of 49 images each A total of 6414 Yoruba handwritten character images of 40 classes with image sizes ranging between 7.71 and 13.5kB were used to trained the modified Vision Transformer (ViT) classifier; an average of 160 images each for each character. The second group of volunteers with twenty writers write one word each. The word image sizes range between 10.5 and 13.3kB while image dimension ranges between 74×40 and 231×41 pixels. Twenty word samples with 34 distinct classes (classes of character) varying from three-letter word to ten-letter word were used as testing dataset to evaluate the performance. Figure 4 shows the sample of the handwritten words. These words were extracted from a popular Yoruba text called 'Aláwìíyé' book 1, 2 and 3 sixth edition authored by J. F. Odunjo. The selection of these words for this research was based on the simplicity of the text, used of simplified methods and pictures used by the text to teach beginners both children and adults how to read and write in Yoruba.

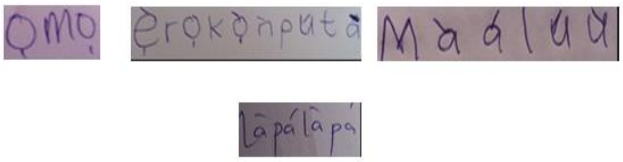

Figure 4: Sample of Yoruba Handwritten word images

Image Preprocessing

The next stage after image acquisition (data acquisition) is image preprocessing stage. Image preprocessing techniques are applied to digital image. This step is necessary in image processing so as to obtain better features and high recognition accuracy. The image pre-processing techniques that were used for this work include image resizing, noise removal and grayscale conversion to improve the quality of the input image.

Image Resizing

The RGB input images in the datasets were resized to 72× 72 pixels to have a uniform image. This was necessary due to

square shape (fixed size) of dataset used for Vision Transformer (ViT) model. The model Split an image into non-overlap patches of equal sizes.

Grayscale Conversion

This is one of the image preprocessing techniques used in computer vision and pattern recognition. This converts the handwritten text image to grayscale image. Grayscale conversion convert the original RGB images into grayscale images for further processing so as to reduce memory space consumed by RGB images. A grayscale image is a scale of shades from black to white used in image technology. An image is represented based on the number of pixels (dimension of height and width). Equation 1 gives the equation for grayscale conversion given by Oladele et al. (2020)

$$GY = 0.59G + 0.29R + 0.12B. \quad (1)$$

Noise Removal

Gaussian filter is used to reduce the noise in an image. Gaussian Filter is a linear Filter and a non-uniform low pass Filter which is the result of blurring an image by a Gaussian function. It is also known as image blurring. This filter is used to reduce image noise by blurring an image. The Gaussian filter mathematical equation for two-dimensional Gaussian function used is given by (Kaur and Singh, 2017; Sann et al., 2021) as:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

Image Segmentation

The next stage after image preprocessing is image segmentation. To achieve this, Otsu thresholding and bounding boxes is used to separate background from foreground and to separate characters from words.

Binarization/Image Thresholding

Binarization is a form of image segmentation in which an image is divided into constituent objects. Binarization also known as Image thresholding is the conversion of grayscale image into black and white (binary image). Otsu thresholding technique is a segmentation technique used to separate foreground from background of an image. Otsu thresholding is applied to the output of the preprocessed image which is a grayscale image. Figure 5 shows a picture of binarized image. Equation for binarization is given by Kaur and Kaur (2014) as;

$$g(x,y) = f(x) = \begin{cases} 1 & \text{if } f(x,y) \geq T \\ 0 & \end{cases} \quad (3)$$

where T is the threshold value.

The threshold value is calculated to reduce interclass variance between black and white pixels. The number of pixels for any gray level i is denoted n_i and normalized as p_i where $p_i > 0$, $\sum_{i=1}^{L} p_I = 1$. To find the optimal threshold, its initial value k is set to zero and repeatedly set to the values for all gray levels up to L. The result of this algorithm classify image into foreground and background. The image is assumed to contain two classes, one for background and the other for foreground (object). The highest gray level found in the image is denoted $1 \leq L \leq 255$. The probability for each of these classes (Kaur and Kaur, 2014) is denoted by:

$$w_0 = \sum_{i=1}^{k} p_I \quad (4)$$

$$w_i = \sum_{i=K+1}^{L} p_I \quad (5)$$

Class variances are then calculated using

$$\sigma_b^2(t) = w_0(t)w_1(t)[\mu_1(t) - \mu_0(t)]^2 \quad (6)$$

where μ is class mean and μ_i is a mean of class i

The threshold value obtained is applied to all the images in the dataset to separate the foreground from background.

Figure 5: Picture of Binarized Handwritten image

Bounding Boxes

The preprocessed image is segmented using character segmentation into individual character for easy recognition. Segmentation will help the classifier to extract features from each individual character. Character segmentation was done using bounding boxes as shown in Figure 6.

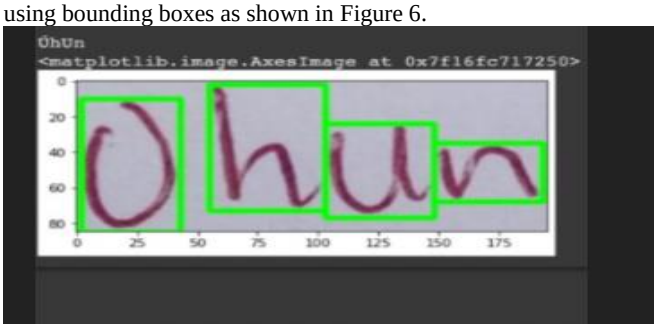

Figure 6: Character segmentation using bounding box

Word Recognition using Modified Vision Transformer

After the completion of image segmentation, the next stage is classification. Vision Transformers model was used for classification of different class of Yoruba handwritten characters. Vision Transformer divides an image into a sequence of flattened image patches and passing the sequence through the transformer model, taking the first element of the transformer's output sequence and passing it through a final classification model (Dosovitskiy et al., 2021). The architecture of ViT model is given in table 1.

Table 1: Vision Transformer Algorithm

Algorithm: Vision Transformer Algorithm (Adapted from Khawar (2022))
Step 1: Split an image into non-overlap/overlap patches (144 patches each with fixed sizes of 6x6)
Step 2: Flattening of the image patches
Step 3: Creation of lower-dimensional linear embeddings from flattened image patches
Step 4: Inclusion/addition of positional embeddings
Step 5: Feeding the linearly projected flattened patches as an input to the encoder

Step 6: Pre-training the ViT model.

Step 7: Image classification

Vision transformer parameter selection and tuning was used to obtained trainable parameters. The parameters used were image shape of height, width and channels (height (p), width (p) and channel (c)) (72×72×1; 32x32x1), patch size of 6 (6×6), four Multi-Layer Perceptron (MLP) heads, six and eight transformer layers. The parameter tuned were image size and number of transformer layers while patch size and number of heads have fixed values in this research. Image sizes used were 32x32x1and 72x72x1 while 6 and 8 number of transformer layers was used.

Eight transformer layers, image shape of 72×72×1, patch size of 6 and four Multi-Layer Perceptron heads with 21,692,904 trainable parameters were obtained from initial empirical experiments. Figures 7 and 8 show the framework of Vision Transformer and flowchart of developed Word-Level Yoruba Handwritten Character Recognition System respectively.

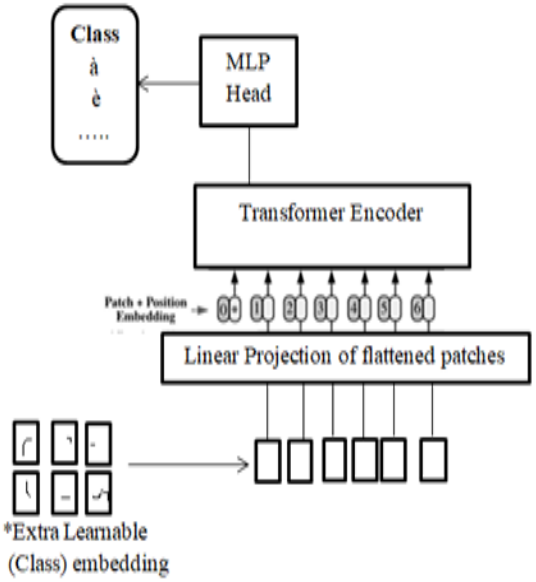

Figure 7: The framework of Vision Transformer (Adapted from Dosovitskiy et al., 2021)

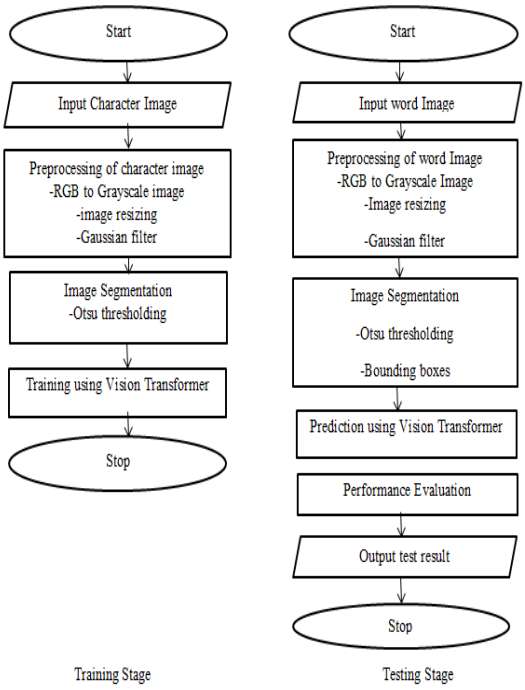

Figure 8: Flowchart of developed Word-Level Yoruba Handwritten Character Recognition System (Training and Testing Stages)

Performance Evaluation

This study employed hold-out evaluation method where the character dataset was used for training and validation while the word-level dataset was used for testing the system for improved accuracy. Furthermore, the study was then evaluated using evaluation metrics like accuracy and testing time.

(i) Accuracy: This measures the overall effectiveness of the developed system and it is measured in percentage (%). Two types of accuracy were used to evaluate the performance of the system. They are;

(a) Segmentation accuracy: This measures the segmentation accuracy of the system in terms of how the words were segmented into individual character. It is measured in percentage (%) and given as;

$$\text{Segmentation Accuracy} = \frac{Number\ of\ Correcly\ segmented\ characters}{Total\ number\ of\ characters\ in\ the\ word} \quad (7)$$

(b) Character-Level Recognition Accuracy: This measures the recognition accuracy of the system in terms of recognition of individual character in a word. It is measured in percentage (%) and given as;

$$\text{Character-Level Recognition Accuracy} = \frac{Number\ of\ correctly\ classified\ characters}{Total\ number\ of\ characters\ in\ the\ word} \quad (8)$$

(ii) Testing time: This determines the average time taken to classify word images by the developed system and it is calculated in seconds (s). Average testing time is calculated by dividing the total testing time with the number of testing dataset.

Training

Training dataset was used for training the model. In training the model, 6414 lower case handwritten images were used as training dataset while 20 lower case handwritten word images were used as testing dataset. The training of the model was conducted using 100 epochs at 1iteration per epoch with total iterations of 100 iterations. The training time was calculated and estimated to be 23minutes 3seconds after 100 epochs. The model yielded network accuracy of 97.2%. Figure 9 shows network performance diagram during model training process while Figure 10 shows loss curve diagram during model training process.

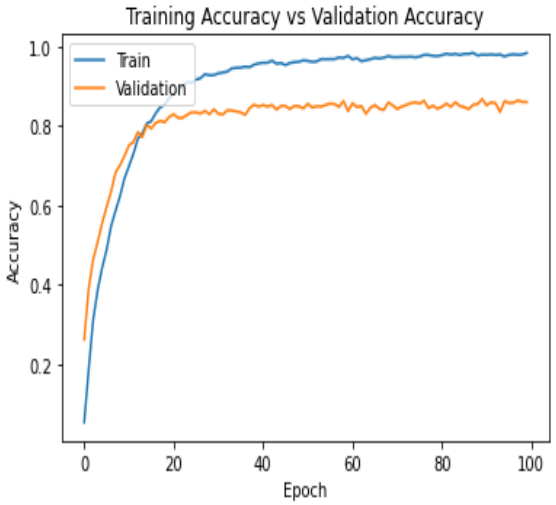

Figure 9: Network Performance diagram during training process

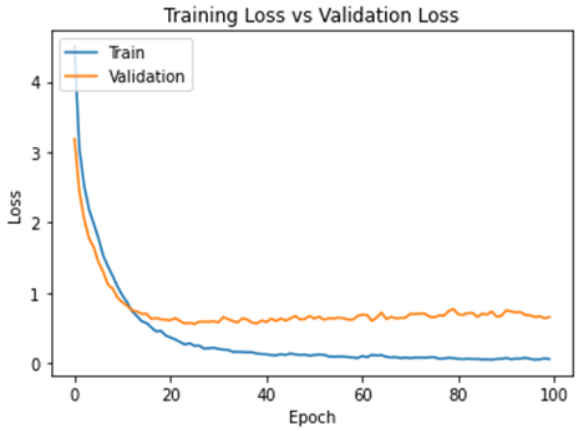

Figure 10: Loss curve diagram during model training process

Figure 9 show that the network accuracy (Network Performance) of the model during training was 0.9720 after 100 epochs. The loss curve diagram in Figure 10 shows that the training loss at final 100 epochs was 0.078 during model training process. It was observed at 80 epochs upward, the curve flatten showing a loss of 0.028 up till 100 epochs.

Results and Discussion

Word-level Yoruba Handwritten Character Recognition (WYHCR) System was tested on a 4GB RAM, Intel core i5, 2.40GHZ CPU HP laptop computer using Graphical Processing Unit (GPU) from Google Colaboratory. Vision transformer parameters of image shape of 72x72x1, 8 layers transformers, and patch size of 6 and 4 heads obtained from initial empirical experiments were used for classification. Twenty Yoruba Handwritten word images were presented to the model as input.

Using ViT parameters of 6 layers transformer, 4 heads, patch size of 6 and image shape of 72×72×1. The performance of the experiment achieved average segmentation accuracy of 79.7%, character-level recognition accuracy of 71.5% and average computational time of 0.9s. Experiment with ViT parameters of 8 layers transformer, 4 heads, patch size of 6 and image shape of 72×72×1, the performance shows average segmentation accuracy of 79.7%, character-level recognition accuracy of 72.1% and testing time of 0.9s. Table 2 shows the performance of ViT parameters on the dataset using image shape of 72×72×1 while Figure 11 shows graphical representation of Performance of ViT parameters on the dataset using image shape of 72×72×1, 6 and 8 layers transformer.

Table 2: Performance of ViT parameters on dataset using image shape of 72×72×1, 6 and 8 layers transformer

Transformer layer	Segmentation accuracy (%)	Character-level recognition accuracy (%)	Testing time (s)
6	79.7	71.5	0.9
8	79.7	72.1	0.9

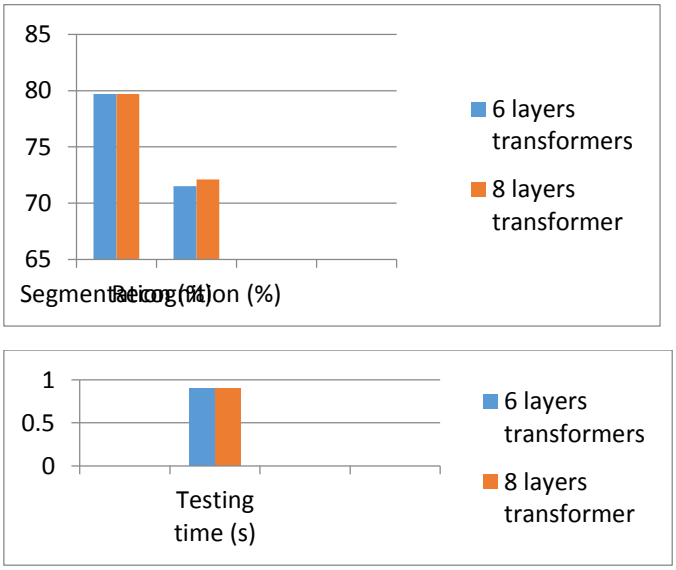

Figure 11: Graphical Representation of Performance of ViT parameters on dataset using image shape of 72×72×1, 6 and 8 layers transformer

The images were reshaped to 32×32×1 and ViT parameters were used to train the model. Dataset with ViT parameters of 6 layers transformer, 4 heads, and patch size of 6 and image shape of 32×32×1 were used. The experiment achieved average segmentation of accuracy of 79.7%, character-level recognition accuracy of 71.4% and testing time of 0.9s. Further experiment with ViT parameters of 8 layers transformer, 4 heads, patch size of 6 and image shape of 32×32×1. The performance of the experiment achieved average segmentation accuracy of 79.7%, character-level recognition accuracy of 71.8% and testing time of 0.9s. Table 3 shows the performance of ViT parameters on the dataset using image shape of

32×32×1 while Figure 12 shows graphical representation of Performance of ViT parameters on the dataset using image shape of 32×32×1, 6 and 8 layers transformer.

Table 3: Performance of ViT parameters on dataset using image shape of 32×32×1, 6 and 8 layers transformer

Transformer Layer	Segmentation accuracy (%)	Character-level recognition accuracy (%)	Testing time (s)
6	79.7	71.4	0.9
8	79.7	71.8	0.9

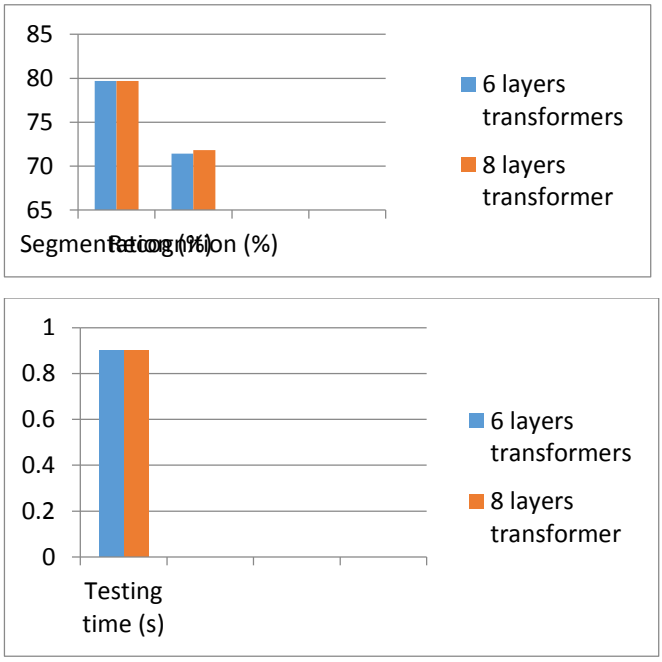

Figure 12: Graphical Representation of Performance of ViT parameters on dataset using image shape of 32×32×1, 6 and 8 layers transformer

The performance of the system achieved an average segmentation accuracy of 79.7%, average character-level recognition accuracy of 72.1% and average testing time of 0.9s. Table 4 shows the performance of the developed WYHCR system with evaluation metrics and testing time while Figure 13 shows graphical representation of Performance evaluation of the developed WYHCR system. From the result, it was observed that 'y' was misclassified as 'ì', 'ò' as 'ọ', 'l' as 'f' and 'h', and 'à' as 'ì'. Character-level accuracy depends on segmentation accuracy. It was also observed that lower segmentation accuracy was due to cursive nature of the words and non-cursive nature of the diacritics with some characters. Low character-level recognition accuracy of some words was due to low segmentation accuracy. Despite moderate character-level accuracy, there were low testing times for each prediction ranging between 0.6s (600ms) and 1.5s (1500ms)

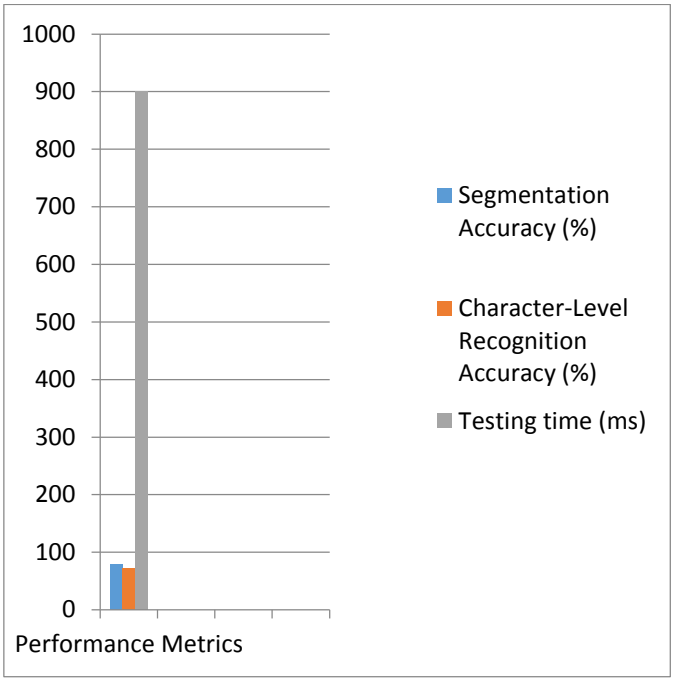

Figure 13: Graphical representation of Performance evaluation of the developed WYHCR system

Table 4: Performance evaluation of developed WYHCR system

S/N	Tested Word	Recognized Word	Segmentation Accuracy (%)	Character-level Accuracy (%)	Testing time (s)
1	pẹ́ńsùlù	pẹ́ńsùlù	100%	100%	1.2s
2	yorùbá	ìorùbá	100%	83.3%	1.1s
3	Ọlọ́run	Ọlọ́run	100%	100%	1.5s
4	Ṣẹ̀kẹ̀rẹ̀	Ṣẹ̀kẹ̀rẹ̀	100%	100%	1.0s
5	Màálúù	Màálúù	100%	100%	1.0s
6	Owó	Owof	66.7%	66.7%	0.8s
7	ọkọ́	ọkọ́	100%	100%	0.7s
8	òro	òro	100%	100%	0.7s
9	òpè	ọ̀pè	100%	66.7%	0.8s
10	ìrùkẹ̀rẹ̀	ìrùkẹ̀	71.4	71.4%	1.0s
11	kọ́ńsóná ǹtì	kọ́ńsmáǹ tì	80%	80%	1.4s
12	Ife	Ife	100%	100%	0.7s
13	Pupa	mpa	50%	50%	0.8s
14	Ọlọ	Ọfọ	100%	66.7%	0.7s
15	ọmọ	ọmoe	66.7%	66.7%	1.1s
16	ẹ̀rọkọ̀ǹp útà	ẹ̀rọkọ̀ǹpú tì	100%	90%	1.5s
17	làpálàpá	wry	0%	0%	0.7s
18	pátákó	kwkó	33.3%	33.3%	0.8s
19	ìlù	ìhù	100%	66.7%	1.4s
20	ohun	ọw	25%	0%	0.6s

Comparison of Results with Existing Yoruba Handwritten Character Recognition Systems

This research compared the performance of existing character recognition systems with the developed WYHCR system. The developed system was tested on re-implemented Support Vector Machine (SVM) classifier similar to Oladele et al., (2017), it was found that the developed WYHCR achieved better character-level recognition accuracy than SVM with the

dataset used in this research. The developed system was also tested on re-implemented Convolutional Neural Networks (CNN) classifier similar to Oyeniran and Oyebode (2021); it was found that the developed WYHCR system outperformed CNN classifier using the dataset in this research. Table 5 shows the comparison of existing character recognition system with developed WYHCR system while Figure 14 shows graphical representation of Comparison of developed WYHCR system with some existing proposed systems.

Table 5: Comparison of Developed WYHCR System with Some Existing Proposed Systems.

S/N	Dataset	Classifier	Recognition Accuracy	Testing time (s)	Character Recognition System
1	Locally acquired dataset	SVM	58.7%	1.2	Oladele et al. (2017)
2	Locally acquired dataset	CNN	69.4%	1.1	Oyeniran and Oyebode (2021)
3	Locally acquired dataset	Modified Vision Transformer(MVIT)	72.1	1.0	Developed system

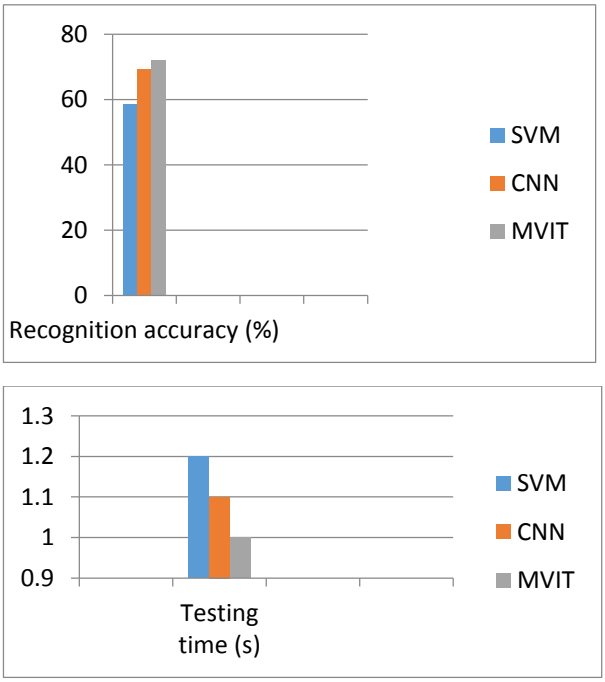

Figure 14: Graphical Representation of Comparison of developed WYHCR system with some existing proposed systems.

Conclusions

The research has developed a word-level Yoruba handwritten character recognition system using modified vision transformer, a deep learning model. The system achieved its aim by developing a word level Yoruba handwritten character recognition using modified vision transformer (ViT) as a classifier. The developed WYHCR system was implemented on Google colaboratory with python programming language. Several parameters of ViT were experimented. The parameters used were image shape, number of transformer layers, number of heads and number of patches. Vision transformer image size and number of transformer layers parameters were tuned to obtained trainable parameters. Findings indicated that with modified Vision Transformer as classifier using image size of 72x72 and transformer with eight layers, the system achieved 21,692,904 trainable parameters with high segmentation accuracy, high character-level accuracy and low testing time. The developed system gave a better performance when evaluated with recognition accuracy and testing time which outperformed SVM and CNN classifiers when compared with existing studies. The developed system performance results indicated that it is a good model for character recognition and could be improved upon and implemented in real time to help digitize Yoruba historical documents and information on paper.

The following are recommended for future work; More research should be done on cursive Yoruba words; Further research should be extended to more complex areas such as sentences, paragraph and document level with improve performance; implementation of Yoruba handwritten character recognition system in real time with live camera and on smart phones for quick access and convenience and other deep learning approach should be implemented for Yoruba words with non-cursive diacritics. Finally, Creation of more Yoruba database for more dataset for training and evaluating the performance of Yoruba Handwritten Character Recognition system.

Acknowledgement

Not applicable

Conflict of Interest

There was no conflict of interest

References

Adeyanju, I.A, Fenwa, O. D., & Omidiora E.O. (2014). Effect of non-image features on recognition of handwritten alpha-numeric characters. *International Journal of Computers and Technology*, *13(11)*, 5155-5161

Adeyanju, I.A., Ojo, S.O., & Omidiora, E.O. (2016). Recognition of typewritten characters using Hidden Markov Models. *British Journal of Mathematics and Computer Science*, *12(4)*, 1-9.

Ajao, J. F., Olawuyi, D. O., & Odejobi, O. O. (2018) .Yoruba handwritten character recognition using freeman chain code and k-nearest neighbor classifier. *Jurnal Teknologi dan Sistem Komputer*, *6(2)*, 129-134.

Awel, M. A., & Abidi, A.I. (2019). Review on optical character recognition. *International Research Journal of Engineering and Technology*, *6(6)*, 3666- 3669

Bo, K.R., Hong, H.S & Wen, H.C. (2022). Vision Transformers: State of the art and research Challenges. *arXiv: 2207.03041v1[cs.CV] 2022*.

Ding, K., Jin, L., & Gao, X. (2009). A new method for rotation free online unconstrained handwritten chinese word recognition. 10th International Conference on Document Analysis and Recognition, 1131-1135

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

Geetha, V. V, Sudheer, A. V., Saikumar & Gomathy, C. K. (2022). Optical character recognition. *Journal of Engineering, Computing and Architecture*, *12(3)*, 211-214.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z.,Zhang, Y. & Tao, D. (2022). A Survey on Vision Transformer. A submission to IEEE transaction on pattern analysis and machine intelligence. arXiv:2012.12556v5 [cs.CV] 23 Feb 2022

Kala, S. L. (2022). Handwritten character recognition using neural networks. *International Journal of Advanced Research in Computer and Communication Engineering*, *11(4)*, 168-171

Kaur, D. & Kaur, Y. (2014). Various image segmentation techniques: A review. *International Journal of Computer Science and mobile computing*, *3(5)*, 809-814.

Kaur, R. & Singh, R. (2017). Image filtering techniques- A review. *International Journal of Advanced research in Science & Engineering*, *6(8)*, 2066-2071

Khawar, I. (2022). Advances in Vision Transformer: A Survey and outlook of recent work. Retrieved on November 20, 2022 from https://github.com/Khawar512/ViT-Survey

Kumar, V. V, Srikrishna, A., Babu, B. R., & Mani, M. R. (2010). Classification and recognition of handwritten digits by using mathematical morphology. *Indian Academy of Sciences*, *35(4)*, 419–426

Nasien, D., Haron, H., & Yuhaniz, S.S (2010). Support Vector Machine (SVM) for English handwritten character recognition. Second International Conference on Computer Engineering and Applications, 249-252

Odunjo, J. F. (2017). Alawiiye, 6th edition. Learn African Plc. Book 5. Page 7. ISBN 978 978 925 396 8

Ojumah, S., Misra, S., & Adewumi, A. (2018). A database for handwritten yoruba characters. In Communications in Computer and Information Science. doi: 10.1007/978-981-108527-7_10

Oladele, M., Adepoju, T.M, Olatoke, O.A., & Ojo, O.A. (2020). Offline yoruba handwritten word recognition using geometric feature extraction and support vector machine classifier. *Malaysian Journal of Computing*, *5 (2)*, 504-514

Omidiora, E.O, Adeyanju, I.A., & Fenwa, O.D. (2013).Comparison of machine learning classifiers for recognition of Online and offline handwritten digits. *Journal of Computer Engineering and intelligent systems*, *4(13)*, 39-48.

Oyeniran, O., & Oyebode, E. (2021). Transfer learning based offline yorùbá handwritten character recognition system. *Journal of Engineering Studies and Research*, *2(27)*, 89-95.

Oyeniran, O.A., & Oyebode, E.O. (2021). Yorùbánet: A deep convolutional neural network design for yorùbá alphabets recognition. *International Journal of Engineering Applied Sciences and Technology*, *5(11)*, 57-61

Pareek, J., Singhania, D., Kumari, R. R., & Purohit, S. P (2020). Gujarati handwritten character recognition from text image. *Procedia Computer Science,171(2020)*, 514-523

Patil, S., Varadarajan, V., Mahadevkar, S., Athawade, R., Maheshwari, L., Kumbhare, S., Garg, Y., Dharrao, D., Kamat, P., & Kotecha, K. (2022). Enhancing optical character recognition on images with mixed-text using semantic segmentation. *Journal of Sensor and Actual Networks*, *11(4)*, 63.

Salman, K., Muzammal, N., Munawar, H., Syed, W.Z, Fahad, S.K & Mubarak, S. (2022). Transformers in Vision: A Survey. *arXiv:2101.01169v5[cs.CV]2022.*

Sann, S. S, Win, S. S., & Thant, Z. M. (2021). An Analysis of various image preprocessing techniques in butterfly image. *International Journal of Advance Research and Development*, *6(1)*, 1-4

Shetty, J.R., & Heraje, K.N. (2017). Recognition of formatted text using machine learning technique. *American Journal of Intelligent Systems*, *7(3)*, 64-67. DOI:10.5923/j.ajis.20170703.05

Sonara, H., & Pandi, S.G. (2021). Handwritten character recognition using convolutional neural network. In 3rd International Conference on Communication and Information Processing, 1-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, L. (2017). Attention is all you need. *In NIPS, 2017*.