Hausa Character Recognition Using Logistic Regression

Akinbowale Nathaniel Babatunde, Roseline Oluwaseun Ogundokun, Ebunayo Rachael Jimoh, Sanjay Misra, and Deepak Singh

1 Introduction

Handwritten character recognition is a critical topic in identifying characters, and it may be used as an exam instance for concepts such as pattern recognition and machine learning's high-efficiency models. Many common elements have evolved to aid machine learning and pattern recognition research [1, 2]. Automatic text extraction from images or videos is a critical and difficult topic in a variety of applications, including document processing, image indexing, video content summary, video retrieval, and video comprehension [3]. Unlike character recognition for scanned documents, recognizing characters in unconstrained images is confounded by a wide scope of variation in backgrounds which consist of irregularities in quality content and detachment of difficult ink from the background [4].

Handwriting recognition (HWR), also known as handwritten text recognition, is a computer's capacity to recognize and understand comprehensible handwritten input from sources such as paper documents, pictures, touch displays, and other

A. N. Babatunde
Department of Computer Science, Kwara State University, Malete, Kwara State, Nigeria
e-mail: akinbowale.babatunde@kwasu.edu.ng

R. O. Ogundokun
Department of Multimedia Engineering, Kaunas University of Technology, Kaunas, Lithuania
e-mail: rosogu@ktu.lt; ogundokun.roseline@lmu.edu.ng

E. R. Jimoh
Department of Computer Science, University of Ilorin, Ilorin, Nigeria

S. Misra
Department of Computer Science and Communication, Ostfold University College, Halden, Norway

D. Singh (✉)
National Institute of Technology Raipur, Raipur, India
e-mail: dsdsingh.cs@nitrr.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

D. S. Sisodia et al. (eds.), *Machine Intelligence Techniques for Data Analysis and Signal Processing*, Lecture Notes in Electrical Engineering 997,
https://doi.org/10.1007/978-981-99-0085-5_65

devices [5]. In the field of HWR, there are two major problem areas. The online mode comprises the automated transformation of the collected text as it is created on a particular digitizer or personal digital assistant, where a sensor takes up pen-tip motions including pen-up or pen-down switching. The collected handwritten text is transformed into letter codes, which are then utilized by text-processing software. The dynamic motion while handwriting is recognized by online character recognition software.

In the offline technique, it involves the system to automatically distinguish and change over text in the form of an image into letter codes; the data obtained are the fixed static shape of the character. In offline HWR (OHR), information apprehended is accessible on paper, which is digitized using a scanner. OHR is relatively hard to comprehend as various individuals have diverse handwriting styles [4].

The Hausa language is one of the utmost frequently articulated innate languages in West and Central Africa, with roughly forty to fifty million persons using it as a primary or alternative language. Within the Afro-Asian language phylum, it belongs to the Western branch of the Chadic language superfamily. The Hausa people's homelands are on both sides of the Niger-Niger border, where roughly half of the people voices Hausa as a primary accent, and Nigeria, where around one-fifth of the people voice it as a primary dialectal [5]. The Hausa people are mostly Muslim. Their long-distance trade and pilgrimages to the Islamic Holy Metropolises have spread their language to almost all key metropolitans in West, North, Central, and Northeast Africa. Subject, verb, and object is the fundamental word order. Hausa is a tone language, which means that pitch contrasts matter just as much as consonants and vowels in determining the meaning of a phrase. In Hausa orthography, the tone is not indicated. Accent marks in academic Hausa transcriptions denote tone, which might be high, mediocre, or descending (diacritic).

Despite extensive study on Latin languages, just a few studies on indigenous languages have been published [6]. HWR researchers have looked at recognition in multiple languages as a separate challenge. Each language has its unique collection of characters and properties, making it challenging to create a universal algorithm that works across all languages. Because of its simplicity and low memory requirements, researchers have utilized logistic regression (LR) to extract the properties of letters or numbers. Various aspects of pattern recognition, notably in the area of character recognition, have been successfully used using LR. For character recognition, some researchers have used traditional approaches such as segmentation and recognition, while others have used artificial neural network (ANN)-based methodologies [7]. LR has become an applicable aspect in the field of artificial learning. It's explicit as a statistical analysis model utilized for the prediction of dependent data value by analyzing based on the past perception of a dataset. This methodology empowers an algorithm being used in a machine learning application to classify incoming data based on historical data.

The number of attempts of researchers has improved the efficiency of document image retrieval. Also, research is going on various local languages [5, 8–10]. This work portrays Hausa character recognition; numerous researchers have worked on character recognition of various languages such as English text, Chinese text, and

Yoruba language, but none or very few researchers have worked on Hausa language character recognition, which also prompted taking this research work of Hausa language character recognition. The research tends to introduce the use of deep learning techniques to recognize Hausa text documents.

The remaining part of the article is structured as follows: Sect. 2 presents the methodology and materials used for the implementation of the study. The results and discussion are presented in Sect. 3. The study was concluded in Sect. 4.

2 Materials and Methodology

Before a computer can detect handwritten writing, many processes must be completed. Sample data collection, preprocessing, feature extraction, categorization, and recognition are only a few of them. The flow design for offline handwritten character recognition is shown in Fig. 1. This project's primary focus is on categorization and recognition. The groundwork for implementing a Hausa text character recognition system starts with a collection of diverse Hausa handwritten characters from various people. The character pictures were acquired and recorded in the database using an HP scanner at a resolution of 300 dpi. After preparing the scanned data, a manual categorization method was used to determine each character's intended output. Figure 2 shows the scanned Hausa handwritten characters.

Image preprocessing for optical character recognition assists with setting up an image for subsequent processing (Fig. 1). It helps in expanding the recognition accuracy by eliminating undesirable information that would have been extricated as features. To remove the isolated pixels, the picture is first transformed into a monochrome image, and then the isolated pixels are simply removed from the resulting monochrome image. Preprocessing includes the following stages: Image annotation, grayscale conversion, image binarization, image normalization, and image resizing.

Image segmentation involves separating the text in the image into blocks, lines, words, and ultimately into characters. It is one of the most important stages in this work (Fig. 1). It helps the classifier with extricating highlights from each character. A boundary was set on the character and cropped accordingly such that only the handwritten image was left with the region which contains the necessary features needed for feature extraction. Cropping helps to remove unnecessary features such as the background, thereby leaving just the relevant features.

Feature extraction was used to extract properties that can distinguish a character uniquely and also to remove the properties that can differentiate between similar

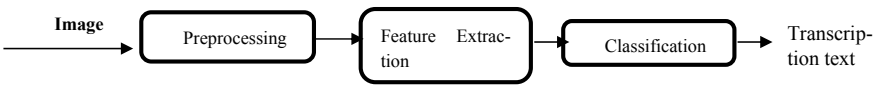


Fig. 1 Block flow of the offline handwritten character recognition



Fig. 2 Samples of captured handwritten of Hausa character

characters (Fig. 1). It distinguishes an area corresponding to a letter from an area corresponding to other letters. It is an essential stage in handwriting character recognition as its effective functioning improves the recognition rate and reduces misclassification. It is utilized to excerpt the features of individual characters which is used to train the system. Statistical feature extraction technique was utilized to excerpt the characters. The major statistical features used in this work are zoning, projections and profiles, and crossings and distances.

The training and recognition stage is the last stage for the developed system. The study used a 10-fold validation before the testing stage was executed. The training and recognition are done using LR. Some of Hausa's uppercase characters will be used for training and testing. For the recognition stage, the feature vector obtained from the feature extraction stage was used for recognition. The feature vectors of the input handwritten image will be compared with the stored feature vectors of the training images to get the result. The unicode of the characters will be used to get the result as the standard keyboard does not have the Hausa alphabet.

Input: Training data

1. For i←1 to k
2. For each training data instance d_i:
3. Set the target value for the regression to
$$z_i \leftarrow \frac{y_j - P(1 \mid d_j)}{[P(1 \mid d_j) . (1 - P(1 \mid d_j))]}$$
4. nitialize the weight of instance d_j to $P(1 \mid d_j). (1 - P(1 \mid d_j)$
5. finalize a f(j) to the data with class value (z_j) & weights (wj)

Classification Label Decision

6. Assign (class label:1) if $P(1|d_j) > 0.5$, otherwise (class label: 2)

Fig. 3 Logistic regression pseudocode

The method used to train the classification of LR is the back-propagation method, and the pseudocode is shown in Fig. 3.

The flowchart for the proposed Hausa character recognition system is shown in Fig. 4. It displays the sequence steps of how scanned images are being recognized from the point when the application is loaded to sending the images into the application via scanner, to the point at which handwriting images are being recognized.

3 Results and Discussion

The designs obtainable and conferred in the previous segment were developed using Octave programming language using C-sharp (C#) programming language to develop the system's user interface to enable users to use the application without accessing the underline model. The desktop application interface relied on the compiled model that was generated by the LR algorithm written in MATLAB programming language. The program was compiled and tested to avoid errors and an installable execution file was generated so that the package can be installed and run efficiently on computer devices regardless of the version of their operating system as long as it meets the necessary hardware requirements. The application was subjected to various testing, and sample

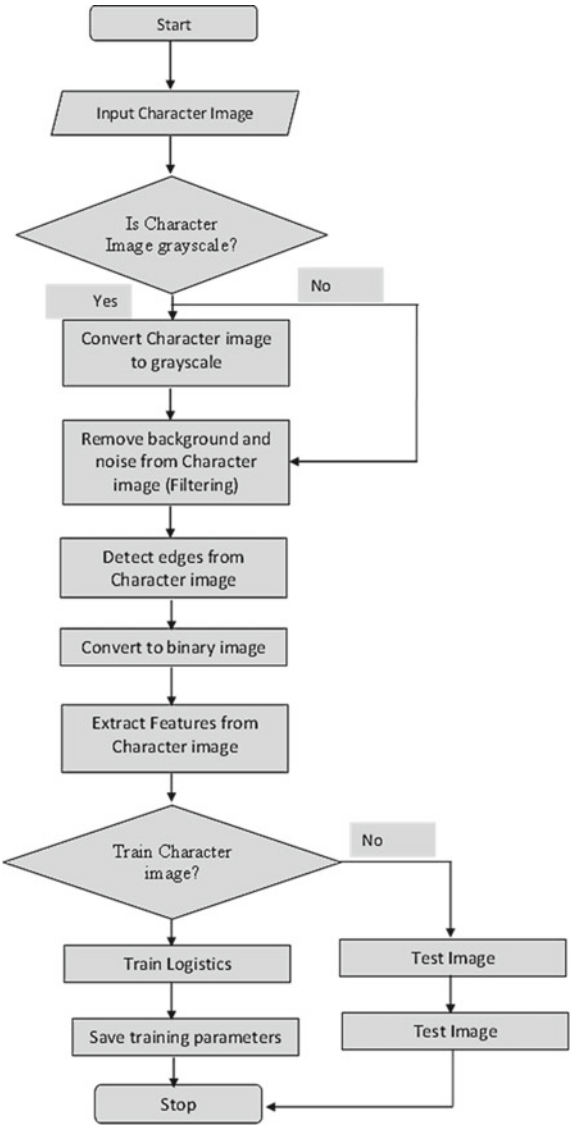


Fig. 4 Flowchart of the proposed system

screenshots were presented. The handwritten character pictures were taken from the constructed Hausa database and tested on the recognition module to verify the built Hausa handwritten character recognition system. Preprocessing was done on the test data. The preprocessed pictures' features were extracted and compared to the learned features vector. If the two feature sets are comparable, it searches the look-up database for the suitable character and displays the matching digital counterpart of the handwritten character if it is found, or returns a false negative if it is not.

4 Results and Discussion

The designs obtainable and conferred in the previous segment were developed using Octave programming language using C-sharp (C#) programming language to develop the system's user interface to enable users to use the application without accessing the underline model. The desktop application interface relied on the compiled model that was generated by the LR algorithm written in MATLAB programming language. The program was compiled and tested to avoid errors and an installable execution file was generated so that the package can be installed and run efficiently on computer devices regardless of the version of their operating system as long as it meets the necessary hardware requirements. The application was subjected to various testing, and sample screenshots were presented. The handwritten character pictures were taken from the constructed Hausa database and tested on the recognition module to verify the built Hausa handwritten character recognition system. Preprocessing was done on the test data. The preprocessed pictures' features were extracted and compared to the learned features vector. If the two feature sets are comparable, it searches the look-up database for the suitable character and displays the matching digital counterpart of the handwritten character if it is found, or returns a false negative if it is not.

The extracted pixel value (0–255) in arrays was visualized in 20×20 sized square; 100 characters were generated randomly from the dataset as shown in Fig. 5. All samples for training and testing are converted to portable network graphics (.PNG) to embellish the improved performance of the proposed scheme.

Samples of handwriting Hausa are displayed in Fig. 5. After getting the binarized image, the designed application extracted each handwritten character pixel within the region of interest module and saved its pixel values that range between 0 and 255.

Extracted handwritten characters are shown in Fig. 6. In this research, a total of 1700 data points from 68 individual writers are used in the training of the model. Each Hausa alphabet character was written with a black ball pen on white background paper in a tabulated form.

The performance evaluation metric used is accuracy, and the accuracy result for the training stage indicated 89%. However, the evaluation result shows an accuracy of about 54%. A built-in Octave function, **fminunc** was used as an optimization solver to find the minimum of an unconstrained function. Optimize cost function J(⍵)with parameters ⍵in LR was used to optimize the algorithm.

The result of testing the implemented model is shown in Figs. 7 and 8 and 9 shows the simple Hausa character recognition with LR model.

Table 1 shows the comparison of the proposed system with existing systems. Several researchers have done different language recognition using machine learning approaches but only three were used for this study. A comparative analysis is shown in Table 1. It was deduced that Desai [11] had an accuracy of 82% using feed-forward back-propagation neural network method, Khemiri et al. [12] had an accuracy of 79% using Naïve Bayes approach, and lastly Jayech et al. [13] had an accuracy of 82% using dynamic hierarchical Bayesian networks. It was shown that our

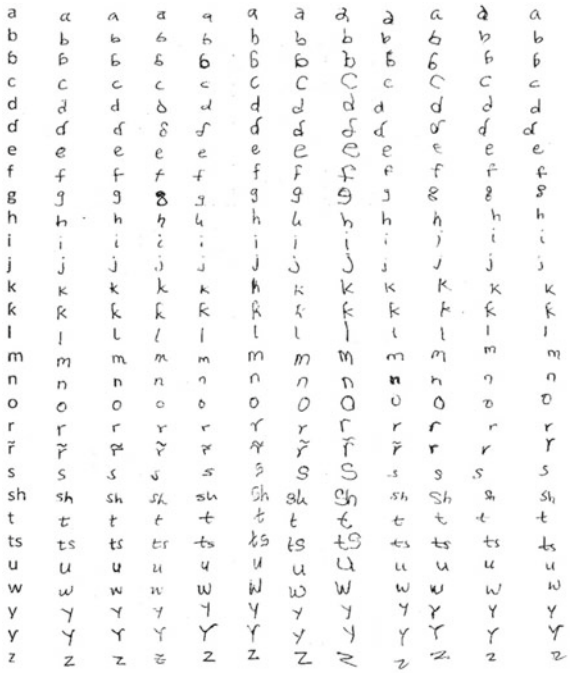


Fig. 5 Samples of Hausa handwriting





Fig. 6 Extracted handwritten characters
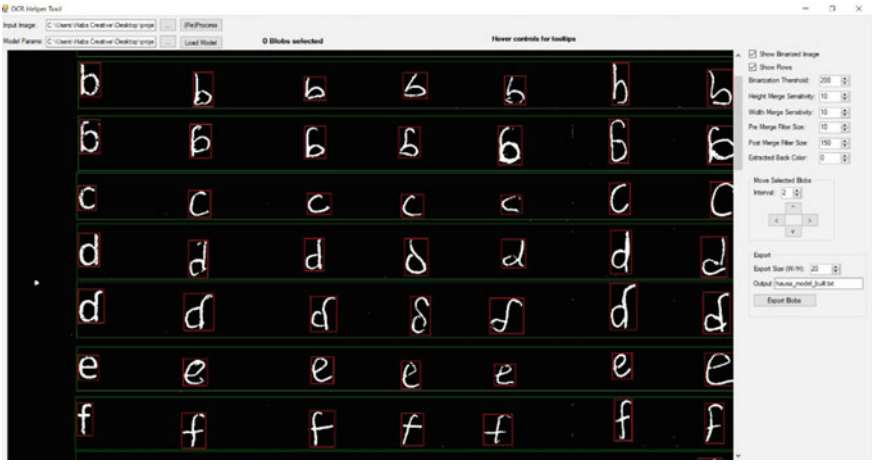


Fig. 7 Character binarization

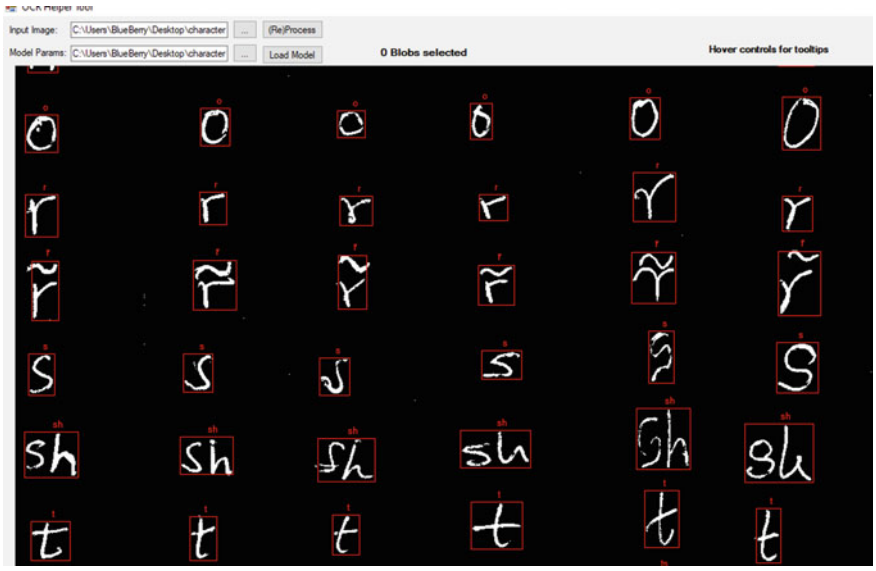


Fig. 8 Character recognition

proposed system is superior by outperforming the three existing systems used for the comparative analysis with an accuracy of 89% (Table 1).

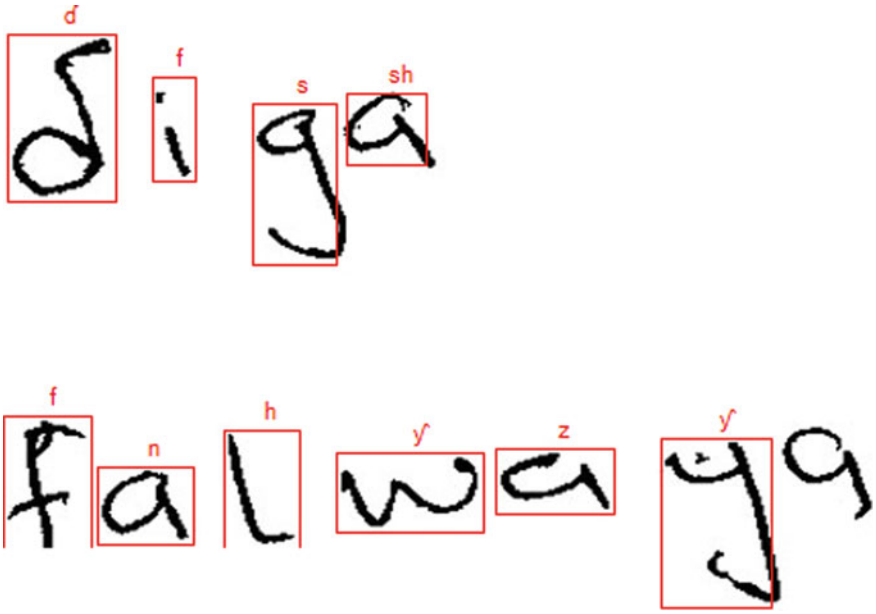


Fig. 9 Simple Hausa character recognition with logistic regression model

Table 1 Comparative analysis with existing systems

Authors	Methods	Accuracy (%)
Desai [11]	Feed-forward back-propagation neural network (FFBPNN)	82
Khémiri et al. [12]	Naïve Bayes (NB)	79
Jayech et al. [13]	Dynamic hierarchical Bayesian networks (DHBN)	82
Proposed system	Logistic regression	89

5 Conclusion

The distinct factor of this work which differentiates this from other related ones is that it has not been implemented using LR. However, a few researchers have tried using the later algorithm, yet they have not been able to accomplish up to 70%of what the project is to accomplish. Some algorithms that have been implemented in accomplishing a similar objective include ANN, K-nearest neighbors, and support vector machine algorithm.

For this research study, the research uses Hausa characters. The previous project has drawn in the utilization of other languages and other classification algorithms. In any case, this research focuses on the use of the Hausa character. The system's user interface was developed with C-sharp (C#) programming language. The desktop

application interface relied on the compiled model that was generated by the LR algorithm written in Octave programming language.

The limitation of the proposed system is dataset collection. As the study deals with language recognition and the language is not a universal one, so getting the Hausa language datasets for the implementation of the study was very difficult. Therefore, more dataset generation is suggested for future work and similarly, more machine learning algorithms or deep learning algorithms can be used in future to have a high accuracy and robust system.

References

1. Gupta A, Sarkhel R, Das N, Kundu M (2019) Multiobjective optimization for recognition of isolated handwritten Indic scripts. Pattern Recognit Lett 128:318–325
2. Blessing G, Azeta A, Misra S, Chigozie F, Ahuja R (2019) A machine learning prediction of automatic text-based assessment for open and distance learning: a review. In: International conference on innovations in bio-inspired computing and applications. Springer, Cham, pp 369–380
3. Jain K, Yu B (1998) Automatic text location in images and video frames. Pattern Recogn 31(12):2055–2076
4. Ajao FJ, Olawuyi OD, Odejobi OO (2018) Yoruba handwritten character recognition using freeman chain Code and K-Nearest Classifier. Jurnal Teknologi dan Sistem Komputer 6(2):129–134. https://doi.org/10.14710/jtsiskom.6.4.2018.129-134
5. Ojumah S, Misra S, Adewumi A (2017) A database for handwritten Yoruba characters. In: International conference on recent developments in science, engineering, and technology. Springer, Singapore, pp 107–115
6. Babatunde AN, Abikoye CO, Oloyede AA, Ogundokun RO, Oke AA, Olawuyi HO (2021) English to Yoruba short message service speech and text translator for android phones. Int J Speech Technol 24(4):979–991
7. Haraty R, Ghaddar C (2004) Arabic text recognition. Int Arab J Inf Technol 1:2
8. Akman I, Bayindir H, Ozleme S, Akin Z, Misra S (2011) A lossless text compression technique using syllable-based morphology. Int Arab J Inf Technol 8(1):66–74
9. Sharma I, Anand S, Goyal R, Misra S (2017) Representing contextual relations with Sanskrit word embeddings. In: International conference on computational science and its applications. Springer, Cham, pp 262–273
10. Ajayi LK, Azeta A, Misra S, Odun-Ayo I, Ajayi PT, Azeta V, Agrawal A (2020) Enhancing the low adoption rate of M-commerce in Nigeria through Yorùbá voice technology. In: International conference on hybrid intelligent systems. Springer, Cham, pp 516–524
11. Desai AA (2010) Gujarati handwritten numeral optical character reorganization through neural network. Pattern Recogn 43(7):2582–2589
12. Jayech K, Trimech N, Mahjoub MA, Amara NEB (2013) Dynamic hierarchical Bayesian network for Arabic handwritten word recognition. In: Fourth international conference on information and communication technology and accessibility (ICTA). IEEE, pp 1–6
13. Khémiri A, Echi AK, Belaïd A, Elloumi M (2016) A system for off-line Arabic handwritten word recognition based on Bayesian approach. In: 2016 15th international conference on frontiers in handwriting recognition (ICFHR). IEEE, pp 560–565