OCR System for the Recognition of Ethiopic Real-Life Documents

Hagos Tesfahun Gebremichael[1(✉)], Tesfahunegn Minwuyelet Mengistu[1], Million Mesheha Beyene[2], and Fikreselam Gared Mengistu[1]

[1] Bahir Dar University, Bahir Dar, Ethiopia
[2] Addis Ababa University, Addis Ababa, Ethiopia

Abstract. A bulk of real-life documents contain vital information and knowledge about history, culture, economy, politics, religion, and science that are written in Ethiopic script. This knowledge has to be shared and the advancement of technology like Optical Character Recognition (OCR) brings the need to digitize documents and make them available for public use. OCR is a process that allows printed, typewritten, and handwritten text to be recognized optically and converted into a machine-readable format that can be accepted by a computer for further processing. Nowadays, effective OCR systems have been developed for languages, like English that has wider use internationally. Researches in the area of Amharic OCR are ongoing since 1997. Attempts were made in adopting recognition algorithms to develop Amharic OCR. This study is, thus, an attempt made to develop an OCR system for real-life documents written in Ethiopic characters. In this study we propose a novel feature extraction schema using Gabor Filter and Principal Component Analysis (PCA), followed by a Genetic Algorithm (GA) based on supported vector machine classifier (SVM). The prototype was tested on real-life Ethiopic documents such as books, newspapers, and magazines, in which an average accuracy of 98.33% for Ethiopic characters is registered.

Keywords: Ethiopic scripts · OCR system · Gabor filter · PCA · GA · SVM

1 Introduction

Over the centuries, paper documents have been the principal instrument to make the progress of humankind permanent [1]. A large number of real-life documents written in Ethiopic script provide essential information and knowledge about history, society, business, politics, religion, and science. Huge collections of documents are archived, written in Ethiopic script in formats such as handwritten, typewritten, or computer printouts [2], which must be translated into electronic form for easy searching and retrieval based on users' information needs or queries. It is sufficient to highlight the massive number of documents in the form of correspondence letters, periodicals, newspapers, pamphlets, and books that are piled up in information centers, libraries, and offices [3]. These documents contain information related to religion, history,

M. L. Berihun (Ed.): ICAST 2021, LNICST 411, pp. 559–574, 2022.
https://doi.org/10.1007/978-3-030-93709-6_38

literature, politics, economics, philosophy, tradition, culture, nature, and other essential shreds of evidence of different nations, nationalities, and people of Ethiopia. Revealing and retrieving the knowledge preserved using Ethiopic script will have a positive impact on social and historical studies. Digitizing this information and allowing public access to these documents will decrease the problems of manual searching and retrieval, and provide vital importance for researchers, historians, tourists, and in general to build a good image about the country and the people. OCR strategies, algorithms, techniques, and tools can be used for this purpose.

Ethiopia is one of the world's ancient countries. It has a well-defined history of more than three thousand years, an ancient and well-developed educational system, philosophy, and writings that are uniquely attributed to it. Furthermore, the country has its language with its alphabets and numerical system, manuscripts, arts, calendars, and hymns which make it unique from all African countries. Most of such identities of the country are found being written in Geez language that is the ancestor of modern Ethion-Semitic languages like Tigrinya and Amharic [2, 3]. Most Ethiopic languages have their own indigenous or native scripts. As a result, libraries, information centers, museums, and businesses have a large number of printed papers. Also, there are plenty of texts which have been circulating among governmental, non-governmental, and private sectors. The digitization of these texts allows existing language technologies to be used to local information demands and advances. The processes by which documents that are created digitally are prepared for electronic access and further research fundamentally from the preservation, searching, and accessing of paper-based texts. Having these types of texts integrated into our digital lives has made OCR such an important technology over the past few years. Due to the trend of IT, and the reasons mentioned earlier, converting Ethiopic documents into electronic format is needed. To convert the text on these documents the conventional way is typing through the keyboard, which is not only time consuming, error-prone, and tedious but also impossible because of the magnitude of documents [4, 5]. The problem of typing into computers is even worse for Ethiopic characters were typing each character needs two keystrokes on average. This emphasizes the importance and tremendous need for developing an OCR system that is capable of recognizing Ethiopic characters. Thus, if automation of documents is needed an OCR software is the preferred means for converting existing documents into machine-readable form.

OCR is a type of document image analysis where a scanned digital image that contains either machine printed or typewritten or handwritten script is input into an OCR software engine and translating into an editable machine-readable digital text format (like ASCII or UNICODE text) [6, 7]. To do this, the OCR system goes through a series of steps, including image acquisition, preprocessing, segmentation, feature extraction, classification, and post-processing, before returning the recognized text documents [8, 9]. By transforming large amounts of printed or handwritten documents into electronic form for subsequent processing, OCR systems make it easier to use computers in everyday life.

2 Related Work

In the 1950s, research and development on automatic character recognition began. Since then, numerous studies on the recognition of various scripts have been conducted, including Latin, Arabic, Chinese, Hindu, Swedish, Russian, Tibetan, Japanese English, Devanagari, Bangla, Farsi, and Kannada, among others. Totally, the complete method is carried out in three phase preprocessing, feature extraction and recognition. In this paper, only Amharic scripture has been studied, which is spoken in countries such as the United States, Israel, Eritrea, Somalia, and Djibouti. Ethiopia is also the only African country with its own indigenous alphabets and writing systems, which is the Geez or Amharic alphabet, while the majority of African countries use English and Arabic scripts or alphabets. [4, 10, 11].

The technique for recognizing Ethiopic characters, on the other hand, is still in its infancy, with the first published study arriving very lately [12]. Furthermore, classifiers face extra challenges due to the structural complexity and interclass similarity of Ethiopic characters. Studies in the application of OCR techniques to Amharic characters have been started in 1997 by Worku et al. [13]. Worku et al. researched the application of OCR techniques to the Amharic characters. His character recognition took into consideration the normal typestyle of WashRa font with a 12-point font size. Under Worku's recommendation, Ermias et al. (1998) conducted more study on the recognition of structured Amharic text as a continuation of Worku's efforts. The goal of this study was to add pre-processing techniques to previously used recognition algorithms so that they could recognize formatted Amharic texts. Ermias used preprocessing techniques for thinning italicized style and underline identification and removal because papers written in Amharic characters occur in diverse font sizes, underline style, and contain italics feature. He incorporated the thinning and underline removal algorithms with the previously adopted recognition algorithm to test the performance of the system [14]. Dereje also attempted to further study in the field in 1999, with the goal of improving Amharic OCR so that it can recognize typewritten Amharic text. Based on his findings, Dereje primarily proposed that in order to improve the Amharic OCR system's recognition accuracy, recognition algorithms that are less sensitive to the peculiarities of character writing styles be used [15].

In the year 2000, Million et al. conducted research in this field with the goal of investigating and extracting the properties of Amharic characters in order to generalize the previously used recognition algorithm to handle diverse Amharic character typefaces [16]. By the same year, Negussie et al. had investigated the recognition of handwritten Amharic legal amounts of bank checks, the purpose of which is to investigate the application of OCR system approaches employed for other characters [17]. In 2002, Yaregal et al. conducted his research in continuation of the research activities done so far to explore the Amharic OCR development approaches, techniques, and methodologies and to come up with a versatile algorithm that is independent of the font size and other quantitative parameters of Amharic characters [4].

Another study by Million et al. was titled Optical Character Recognition of Amharic Documents and was published in 2007. This paper presents an OCR system for converting digitized documents in local languages [18]. Yaregal et al. investigated Hmm-based handwritten Amharic word recognition using feature concatenation in 2009. For offline handwritten text, this paper proposes writer-independent HMM-based Amharic word recognition [19].

Most of the previously done OCR system algorithms for Amharic scripts are based on specific font type, sizes and styles such as Washra, Power Geez, and Visual Geez etc. As this research is a continuous of the previous works, this study focuses on developing generic algorithm and independent of font types, sizes and styles for the recognition of Ethiopic real-life documents using SVM classifier. The contribution of this research is to develop a generic algorithm for Ethiopic real-life documents, to prepare training and testing datasets, we use Gabor Filter for feature extraction and Genetic Algorithm, which has never been used by previous researchers, for selecting the best features of the Ethiopic Characters, to study the experimental result using different kernel functions and analyze which kernel function has scored better result.

3 Ethiopic Character and Proposed OCR System

Ethiopic languages are written from left to right, with no distinction between capital and lowercase letters [16, 17]. As a fundamental character, the Ethiopic writing system consists of 35 characters (named "Fidel"/ "ፊደል"). The thirty-five fundamental characters are organized into seven groups, each of which represents a syllable with a consonant and a vowel after it. Non-basic forms are descended from basic forms through a series of more or less regular alterations. There are also symbols for labialization, numbers, and punctuation signs accessible. With these additions, the script now contains 328 different alphabets, including 245 basic characters, 54 labialize characters, 9 punctuation signs, and 20 numbers [20] (Tables 1, 2 and 3).

The proposed architecture for the OCR System is shown in Fig. 1. Given the digitalized document images of Ethiopic scripts, they are first preprocessed for removing noise, apply binarization, skewness detection, and slant correction. After the image is preprocessed, the next step is segmenting the image into three components: lines, words, and characters. After this stage, the segmented characters are delivered into the feature extraction stage. Under the feature extraction stage, features of the segmented characters are extracted using the Gabor filter. And then the dimensions of the extracted features are reduced using PCA technique. After that, the best features are selected using a GA. Then the best features, selected by GA, are delivered to the classification stage to classify the characters into the correct class using SVM.

Table 1. Ethiopic characters

	Ge'ez a	Ka'eb u	Salis i	Rab'e a	Hamis e	Sadis h	Sab'e o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
sh	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
q	ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
ch	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
gn	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
x	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
x	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
h	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
z	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
j	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
t	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
ch	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
ts	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
ts	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
p	ፘ	ፙ	ፚ	ፗ	፜	፝	፞
v	ቨ	ቩ	ቪ	ቫ	ቬ	ብ	ቮ

Table 2. Labialization characters

ኰ	ኩ	ኵ	ኳ	ኴ
ቈ	ቁ	ቊ	ቋ	ቌ
ቈ	ቁ	ቊ	ቋ	ቌ
ቐ	ቑ	ቒ	ቓ	ቔ
ጐ	ጉ	ጒ	ጓ	ጔ
ኈ	ኍ	ኊ	ኋ	ኌ
ዀ	ዂ	ዄ	ዅ	዆
ጘ	ጙ	ጚ	ጛ	ፘ
ፙ	ፚ	፛	፟	

Table 3. Ethiopic numerals

Ethiopic Numbers	Arabic Numbers	Ethiopic Numbers	Arabic Numbers
፩	1	፳	20
፪	2	፴	30
፫	3	፵	40
፬	4	፶	50
፭	5	፷	60
፮	6	፸	70
፯	7	፹	80
፰	8	፺	90
፱	9	፻	100
፲	10	፼	1000

An SVM classifier uses a well-known algorithm to determine membership in a given class, based on training data. The classifier has two basic functions: training and classification. The process of generating a classifier based on content that is known to belong to specific classes is referred to as training. Classification is the process of using a classifier constructed with such a training content set to assess if unknown information belongs in a specific class.

An SVM training algorithm creates a model that allocates fresh samples to one of two categories given a series of training examples, each tagged as belonging to one. An SVM model is a representation of the examples as points in space, mapped so that the examples of the different categories are separated by a large distance. New examples are then mapped into the same space and classified according to which side of the gap they fall on. Finally, the recognized text documents are obtained.

3.1 Preprocessing Ethiopic Real-Life Documents

The accuracy of the recognition step in OCR systems highly depends on the effectiveness of their preprocessing steps. From the standpoint of this study, the purpose of preprocessing processes is to remove noise and unwanted artifacts from picture data to an acceptable level and provide a refined image for subsequent tasks in character recognition. Thresholding (the task of converting a gray scale image into a binary (black-white) image), noise removal (filtering out background non-textural matters, interfering strokes, shades, and dots introduced by input devices), and skew detection and correction (aligning the paper document with the coordinate system of the scanner) are some of the necessary analyses to perform before recognizing scanned images.

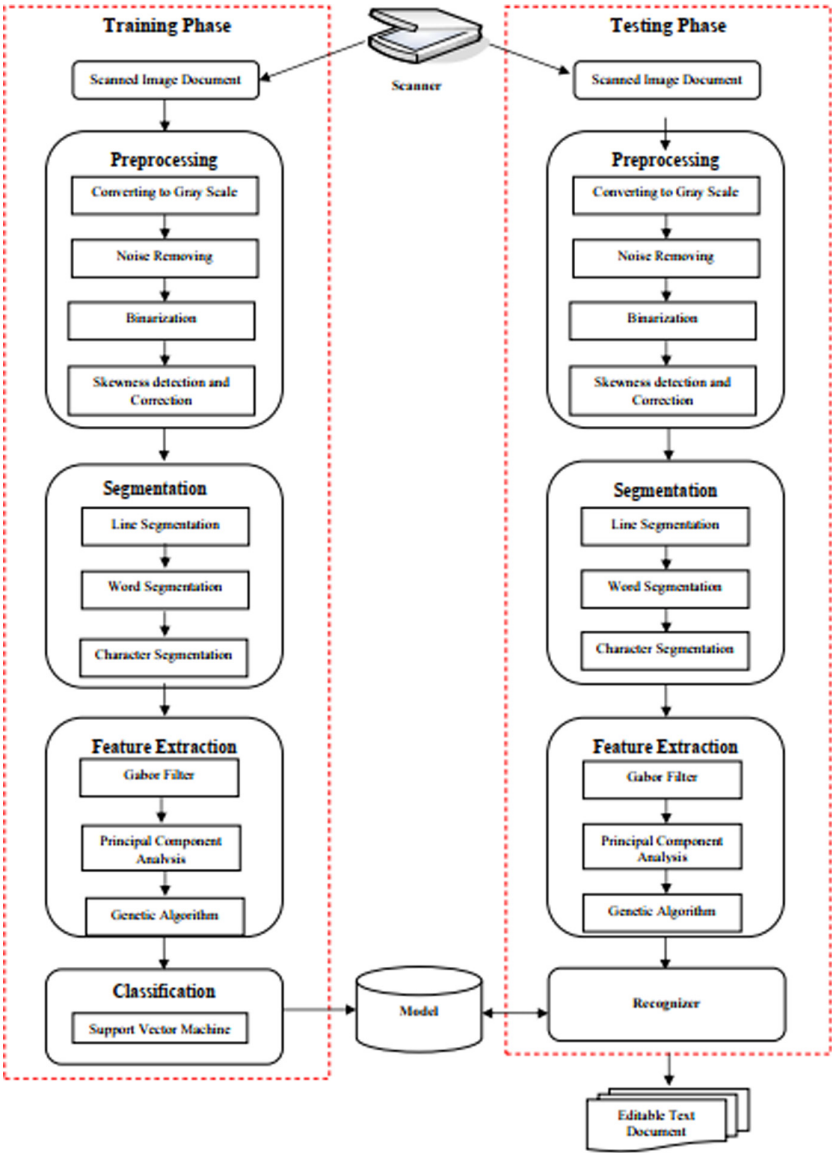

Fig. 1. Architecture of the proposed OCR system for ethiopic real-life documents

Image Segmentation

Line segmentation, word segmentation, and character segmentation are all layers of binary picture segmentation [21].

Line Segmentation: The row histogram was constructed by scanning the input image horizontally and counting the frequency of black pixels in each row to detect text lines.

A boundary between two successive lines is defined as the point where the number of pixels in a row is zero (Fig. 2).

Fig. 2. Line segmentation

Word Segmentation: Following the detection of a line, each line is scanned vertically for word segmentation. A column histogram is made up of the number of black pixels in each column. The section of a line with continuous black pixels is characterized as a word in a line (Fig. 3).

Fig. 3. Word segmentation

Character Segmentation: Each word is scanned vertically for character segmentation once it has been segmented. A column histogram is made up of the number of black pixels in each column. A character in a word is defined as the section of the word with continuous black pixels (Fig. 4).

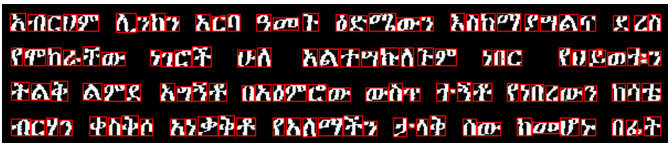

Fig. 4. Character segmentation

3.2 Feature Extraction Steps

The extraction of features is an important aspect of any recognition system. Feature extraction seeks to find patterns with the fewest number of features that are useful in classifying patterns [22]. The performance of the recognition system greatly depends on the features that are being extracted. Using the retrieved features, each character should be able to be classified separately [23]. The Gabor filter was employed to extract features in this study.

1. Gabor Filter

This research presents a new and robust feature extraction method based on Gabor filters for character recognition. Because of their exceptional qualities, Gabor filters are commonly used in image processing and texture analysis: Gabor filters are good for texture representation and discrimination because their frequency and orientation representations are close to those of the human visual system [24].

Gabor wavelets filters have frequency and orientation representations that are close to those of the human visual system, making them suitable for texture representation and discrimination. Gabor filters are widely used in pattern analysis [25]. Gabor filters' most major advantage is its invariance to light, rotation, scaling, and translation. They can also endure photometric disturbances like as variations in illumination and picture noise. A two-dimensional Gabor filter is a Gaussian kernel function modulated by a complex sinusoidal plane wave in the spatial domain, defined as:

$$\begin{aligned} \mathrm{G}\ (x,y) = \frac{f^2}{\pi\gamma\eta} exp\left(\frac{-x^2+\gamma^2y^2}{2\sigma^2}\right) exp(j2\pi fx' + \phi) \\ x' = xcos\theta + ysin\theta \\ y' = -xsin\theta + ycos\theta \end{aligned} \tag{1}$$

where f is the sinusoidal factor's frequency, θ denotes the orientation of a Gabor function's normal to the parallel stripes, ϕ is the phase offset, σ is the Gaussian envelope's standard deviation, and γ is the spatial aspect ratio, which specifies the ellipticity of the Gabor function's support.

As shown in Fig. 5, our proposed algorithm uses forty Gabor filters in five scales and eight orientations to recognize characters, as well as the real parts of the outcome images after applying Gabor filters to the character image. Given the significant correlation between neighboring pixels in the image, we may reduce the information redundancy by down sampling the feature images produced by Gabor filters [24].

Gabor filters extract the character's variations at different frequencies and orientations. The size of the output feature vector is calculated by multiplying the image size (20 × 20) by the number of scales and orientations (5 × 8) divided by the row and column down sampling factors (2 × 2), resulting in a total of 20 × 20 × 5 × 8/(2 × 2) = 4000. Even after down-sampling, the feature vector is still quite huge.

As a result, we'll need to employ dimensionality reduction techniques [26]. PCA was used to reduce dimensionality.

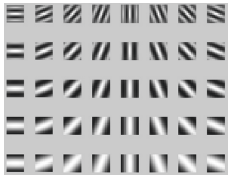

Fig. 5. Gabor filter in five scales and eight orientations.

2. Principal of Component Analysis (PCA)

PCA is one of the methods used for dimensionality reduction. Such a method reduces the data down into its basic components, stripping away any unnecessary parts. This should be done without losing valuable information. Reducing dimensions helps to simplify the data and makes it easier to visualize. For PCA to work, we must subtract the mean from each of the data dimensions. The mean is subtracted from the average of all dimensions. As a result, X' (mean values of all data points) is removed from all X values, and Y' is subtracted from all Y values. This results in a data collection with a mean of zero. Following that, the data's covariance matrix must be calculated. The eigenvector and eigenvalues are computed using this covariance matrix. The eigenvectors and eigenvalues for this matrix can then be calculated. These are critical because they offer us with critical data information. Eigenvectors and eigenvalues are always found in pairs: each eigenvector has an eigenvalue. A direction of the principal component is called an eigenvector. An eigenvalue is a number that indicates the amount of volatility in the data in that direction. The eigenvector with the highest eigenvalue is thus the major component.

3. Genetic Algorithm (GA)

The Gabor filter was used to extract the feature, which was then reduced in dimensionality using PCA before being applied to the Genetic Algorithm (GA) to provide the optimum features for optimal accuracy. GA is a set of population-based and algorithmic search heuristics that replicate the natural evolution process [2, 27]. A population of individual solutions is repeatedly modified by the genetic algorithm. At each phase, the genetic algorithm chooses parents at random from the current population and uses them to produce the following generation's children. Over successive generation the population evolves towards an optimal solution. Based on this the surviving chromosome for GA is the string best chromosome = {1 1 0 1 1 1 1 1 0 0 0 1 0 1 1 1 1 0 1 1 1 1 1 0 1 1 0 0 0 1 1 1 0 1 1 1 1 1 0 0 0 0 0 1 1 0 1 1 1 0 1 0 0 1 0 1 1 1 0 0 0 1 1 1 1 0 1 0 1 1 0 1 0 0 1 0 1 1 1 0 0 0 1 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0}. The positional indices of "1s" in this string are {1 2 4 5 6 7 8 12 14 15 16 17 19 20 21 22 23 25 26 30 31 31 34 35 36 37 38 44 45 47 48 49 51 54 56 57 58 62 63 64 65 67 69 70 72 75 77 78 79 83 85 86 90 91}. The corresponding features from this string are those features positions {1 2 4 5 6 7 8 12 14 15 16 17 19 20 21 22 23 25 26 30 31 31 34 35 36 37 38 44 45 47 48 49 51 54 56 57 58 62 63 64 65 67 69 70 72 75 77 78 79 83 85 86 90 91}, which is 54 features. Equation 2 shows the fitness function which is used in our experiment.

$$\mathrm{F} = \frac{\alpha}{\mathrm{Nf}} + \exp\left(-\frac{1}{\mathrm{Nf}}\right) \tag{2}$$

Where, F, α, and Nf are fitness function, KNN-Based classification error, and Cardinality of the selected features respectively. The algebraic structure of Eq. 2 ensures the learning of the GA, error minimization, and reduced number of features selected.

In GA, we start with the original population and then apply three operators: the first is a selection, which selects the strings, the second is a crossover, which recombines those selected strings, and the third is a mutation, which alters the shape of strings of

0's and 1's. Now we apply the condition and see if the optimization conditions are met. If they are, we choose the best string, which is our solution, and if they are not, we send it to the initial population, as shown in Fig. 6 (Fig. 7).

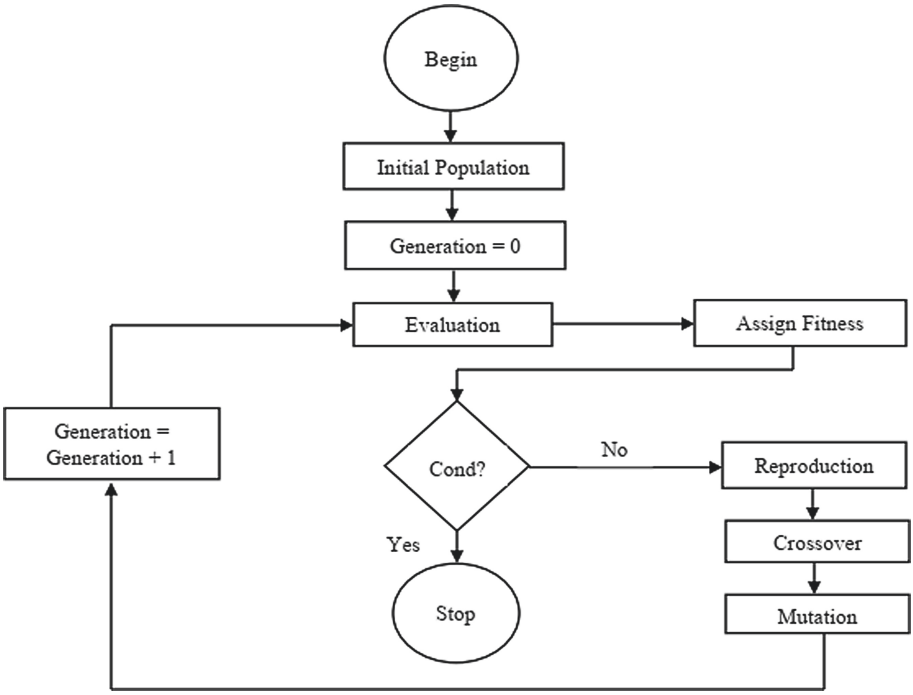

Fig. 6. Flow chart of genetic algorithm.

Genetic Algorithm

Initialization [population];
Evaluation [population];
Generation: = 0;
-do
Selected-parents: = selection[population];
Created-offspring: = recombination[selected-parents];
Mutation [created-offspring];
Population: = created-offspring;
Evaluation [population];
Generation: = geaeration+1;
UNTIL stop-criterion;

Fig. 7. Pseudocode genetic algorithm

4 Result and Discussions

Once the system is developed, an experiment is conducted to evaluate its performance in the recognition of Ethiopic characters. To this end, two experiments are conducted. The first experiment is done using PCA without applying GA and the second experiment is done by applying GA on the reduced feature vectors using PCA. In each experiment, the four kernel functions, such as linear, polynomial, radial basis, and sigmoid of SVM are used to select the best kernel function for the recognition (Fig. 8).

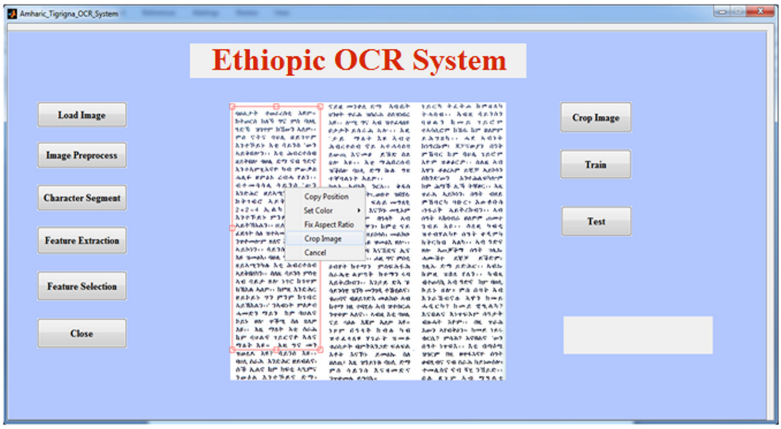

Fig. 8. System prototype interface

4.1 Experiment 1: Performance Analysis Using PCA

In this experiment, first, we extracted 10540 × 4000 features using the Gabor filter from the segmented characters, and then PCA is applied to reduce the extracted features to 10540 × 100 feature spaces. The main reason that we reduced the feature space is that to minimize high memory storage consumption and processor computational time during classification process. In addition to this, as the number of features of characters increase, the similarity of the features among characters should also increase. Hence, this leads to increase misclassification errors during recognition. So, that is why we need to reduce the feature spaces from 4000 to 100 using PCA. After we have reduced the dimensional space of the feature vectors of the datasets into 100 feature spaces, then SVM has been trained with 4540 training data and tested with 6000 testing data. Table 4 shows the performance of the OCR system after applying PCA for feature space reduction.

Table 4. The result of the experiment conducted at 100 features reduced using PCA

Training phase			Testing phase		
Training data	Kernel function	Parameter value	Document type	Testing data	Accuracy (%)
4540	Linear	C = 8	Book	2000	87.8
			Magazine	2000	94.2
			Newspaper	2000	96.8
Average					**92.93**
4540	polynomial	C = 4, γ = 0.25	Book	2000	87.5
			Magazine	2000	95
			Newspaper	2000	96.6
Average					**93.6**
4540	Sigmoid	C = 8, γ = 0.0625	Book	2000	76.2
			Magazine	2000	83.7
			Newspaper	2000	83.6
Average					**81.17**
4540	RBF	C = 8, γ = 0.5	Book	2000	89.4
			Magazine	2000	94.8
			Newspaper	2000	96.7
Average					**93.63**

As can be seen from Table 5, the RBF kernel function registered a better result than the other kernel functions with an average accuracy of 93.63% for Ethiopic characters. From the experimental result, we observed that there is a performance difference among books, magazines, and newspapers. The main reason for performance difference is due to printing variations, character similarity, and document degradation. Document recognition happens because of artifacts such as cuts, merges, ink blobs, etc. that are commonly observed in printed document images scanned from books, magazines, and newspapers. Degradations due to cuts break character components into two or more, and ink-blobs join disjoint characters as one connected component.

4.2 Experiment 2: Performance Analysis Using GA

In this experiment, we extracted 54 features from the character dataset. Thus, the dimensionality of the dataset is reduced to 10540 $\times$ 54. Pattern or image recognition systems may be threatened by high-dimensional feature sets. In other words, having too many features can impair the identification system's classification accuracy because some of them are redundant and non-informative. To attain the best combination of features, several combinatorial sets of features should be obtained. As a result, to limit the number of features used by the SVM Classifier in this study, a GA-based feature selection (a subspace or manifold projection technique) will be applied. A Feature Subset Selection (FSS) is a mapping from an m-dimensional feature space (input space) to an n-dimensional feature space (output space) using the Fs operator:

$$Fs: R^{rm} \rightarrow R^{rxn} \quad (3)$$

where $m \geq n$ and $m, n \in Z+$, $Rr \times m$ is any database or matrix holding the original feature set with r instances or observations, $Rr \times n$ is the reduced feature set including r observations in the subset selection. After selecting the optimal features using GA, the SVM classifier has been trained using these features and the result obtained after applying the GA is shown in Table 5.

Table 5. The result of the experiment conducted at 54 features reduced using GA

Training phase			Testing phase		
Training data	Kernel function	Parameter value	Document type	Testing data	Accuracy (%)
4540	Linear	C = 8	Book	2000	95.7
			Magazine	2000	97.3
			Newspaper	2000	98
Average					**97**
4540	polynomial	C = 8, γ = 2	Book	2000	96.1
			Magazine	2000	97.2
			Newspaper	2000	98.2
Average					**97.17**
4540	Sigmoid	C = 8, γ = 0.0625	Book	2000	84.6
			Magazine	2000	87.4
			Newspaper	2000	88.7
Average					**86.9**
4540	RBF	C = 8, γ = 0.5	Book	2000	97.7
			Magazine	2000	97.9
			Newspaper	2000	98.5
Average					**98.33**

As can be seen from Table 5, the RBF kernel function registered a better result than the other kernel functions. Based on this reason we selected the performance registered by RBF kernel function as accuracy registered by the system. As a result, the system has completed classifying the features of the input characters with an accuracy of 98.33% for Ethiopic characters on average. This shows that an average of 1.67% of Ethiopic characters is wrongly recognized. From this result, we concluded that the accuracy of GA-based dimensionality reduction for feature selection is better than the accuracy of PCA-based dimensionality reduction. The reason why experiment two outperforms better than experiment one is that GA reduces the features of the characters by selecting the most discriminant features that enable the system to recognize correctly some of the similar Ethiopic characters. In this experiment, the performance difference was observed among books, magazines, and newspapers because of printing variations and low document quality. The recognizer correctly classified 97.7% of 2000

characters from books, 97.9% of 2000 characters from magazines, 98.5% of 2000 characters from Newspapers.

In this study, we have used the Support Vector Machine (SVM) for the classification task. The performance of the system is sufficiently promising. Because GA effectively identifies the best properties of a character, the GA-selected features enhanced the classifier's accuracy from 93.63% percent to 98.33% for Ethiopic documents.

The errors encountered in the test result of the developed recognition system can be seen in two broad views: segmentation error and classification/recognition error. Segmentation error is an error that occurred due to the segmentation algorithm. On the other hand, classification/recognition errors occurred in evaluating the performance of the system. The recognition error is an error that occurred due to structurally similarity between the input characters and its recognition result. For example, ሏ is recognized as ሗ.

5 Conclusion

The purpose of this study was to design an OCR system for real-life documents written in Ethiopic languages. The developed system has seven phases: Text digitalization, Preprocessing, Segmentation, Feature extraction, Classification model, Recognizer engine, and finally recognized documents. Finally, the performance of the system has been tested with 6000 unknown instances which are collected from Ethiopic Real-Life documents. The average performance of the correctly classified characters for Ethiopic scripts is 98.33%, which is obtained using GA. The majority of the misclassification errors are attributed to the challenges faced by the segmentation and noise removal techniques from poor quality Ethiopic document images.

One of the major challenges in real-life documents is degradation. So, to enhance the performance of the system there is a need to integrate preprocessing techniques such as advanced noise removal algorithms and also image restoration techniques for blurred images. And also, to improve the performance of the system, an advanced segmentation technique that dynamically adjusts the threshold value according to the document under consideration should be developed.

References

1. Marinai, S., et al.: Introduction to document analysis and recognition. IEEE Trans. PAMI **27** (1), 23–43 (2006)
2. Encyclopedia Britannica, Ethiopia: Encyclopaedia Britannica Ultimate Reference Suite. Encyclopaedia Britannica, Chicago (2010)
3. Meshesha, M., Jawahar, C.V.: Matching word images for content-based retrieval from printed document images. Int. J. Doc. Anal. Recogn. **11**(1), 29–38 (2008)
4. Assabie, Y., et al.: Optical character recognition of Amharic text: an integrated approach (Master thesis). School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia (2002)
5. Belay, B., Habtegebrial, T., Meshesha, M., Liwicki, M., Belay, G., Stricker, D.: Amharic OCR: an end-to-end learning. Appl. Sci. **10**(3), 1117 (2020). https://doi.org/10.3390/app10031117

6. Eikvil, L., et al.: OCR - Optical Character Recognition, pp. 317–326 (December 1993). Norsk Regnesentral, P.B. 114 Blindern, N-0314 Oslo
7. Getu, S., et al.: Ancient Ethiopic Manuscripts Character Recognition, vol. 38 (July 2020)
8. Tanner, S., et al.: Deciding whether optical character recognition is feasible. King's Digital Consultancy Service (December 2004)
9. Ahmed, M., Abidi, A.: Review on optical character recognition. IRJET **06**, 3666–3669 (2019)
10. Cowell, J., Hussain, F.: Amharic character recognition using a fast signature based algorithm. In: Proceedings of the 7th International Conference on Information Visualization, pp. 384–389 (2003)
11. Belay, B., Habtegebrial, T., Belay, G., Meshesha, M.: Learning by injection : attention embedded recurrent neural network for Amharic text-image recognition (October 2020)
12. Alemu, W., et al.: The application of OCR techniques to the Amharic script (Master thesis). School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia (1997)
13. Demilew, F.A., Sekeroglu, B.: Ancient Geez script recognition using deep learning. SN Appl. Sci. **1**(11), 1–7 (2019). https://doi.org/10.1007/s42452-019-1340-4
14. Abebe, E., et al.: Recognition of formatted Amharic text using optical character recognition (Master thesis). School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia (1998)
15. Teferi, D., et al.: Optical character recognition of typewritten Amharic text (Master thesis). School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia (1999)
16. Meshesha, M.: A generalized approach to optical character recognition of Amharic texts (Master thesis). School of Information studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia (2000)
17. Taddesse, N., et al.: Handwritten Amharic text recognition applied to the processing of bank cheques (Master thesis). School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia (2000)
18. Meshesha, M., Jawahar, C.: Optical character recognition of Amharic documents. Afr. J. Inf. Commun. Technol. **3**(2) (2007)
19. Yaregal, A., Josef, B.: HMM-based handwritten Amharic word recognition with feature concatenation. In: Proceedings of the International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009, pp. 961–965 (2009)
20. Trier, O.D., Jain, A.K.: Goal-directed evaluation of binarization methods. IEEE Trans. Pattern Anal. Mach. Intel. **17**(12), 1191–1201 (1995). https://doi.org/10.1109/34.476511
21. Asnake, B., et al.: Retrieval from real-life Amharic document images (Master Thesis). School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia (June 2012)
22. Meshesha, M., et al.: Recognition and retrieval from document image collections. Ph.D. Dissertation, International Institute of Information Technology, India (2008)
23. Mori, M., et al.: Character Recognition. Sciyo (2010). ISBN 978-953-307-105-3
24. Asht, S., Dass, R.: Pattern recognition techniques: a review. Int. J. Comput. Sci. Telecommun. **3**(8), 25–29 (2012)
25. Zhang, D., Lu, G.: A comparative study on shape retrieval using Fourier descriptors with different shape signatures. In: Proceedings of the IEEE International Conference on Multimedia and Expo, Tokyo, Japan, pp. 1139–1142 (2001)
26. Haghighi, M., et al.: Identification using encrypted biometrics. Department of Electrical and Computer Engineering, University of Miami, pp. 440-448 (2013)
27. Akram, S., Dar, M.-U.-D., Quyoum, A.: Document image processing - a review. Int. J. Comput. Appl. **10**(5), 35–40 (2010)