Artificial Neural Network Approach to the Development of OCR for Real Life Amharic Documents

Abay Teshager Birhanu, R. Sethuraman

Lecturer, College of Engineering & Technology, Department of Computing Technology, Aksum University.
Lecturer, College of Engineering & Technology, Department of Computing Technology, Aksum University.

***Abstract*- Although some developments have been made in recognizing various types of machine-printed, typewritten and handwritten Amharic documents, there is a need to enhance its performance on real-life documents which have a number of artifacts that affect the performance of the recognizer. This paper presents the development of Optical Character Recognition (OCR) for real life Amharic degraded documents. For classifying the features generated, an Artificial Neural Network (ANN) approach is implemented.The neural network is trained with eight samples taken from real-life documents. The performance of the developed system is tested with documents taken from real-life documents. Accordingly, an average recognition rate of 96.87% for the test sets from the training sets and 11.40% recognition rate is observed for the new test sets.**

***Keywords*-** Amharic NLP, Optical Character Recognition, ANN Systems

1. INTRODUCTION

OCR is a process that allows printed (typewritten, printout as well as handwritten) text to be recognized optically and converted into machine–readable code that can be accepted by a computer for further processing (Genovese, 1970). OCR systems provide a tremendous opportunity in handling repetitive, boring, labor-intensive, error prone, and time consuming processes for human beings. Postal mail sorting according to destination addresses, bank check processing, bill processing, keying in data to the computer, etc belong to this category. The tasks can be performed with computers in a stable manner (Green, 1993). In order to develop an OCR system it requires the development and integration of many sub systems. The first step is preprocessing such as skew detection and correction, noise detection and removal, binarization, thinning, and normalization. Then segmentation of document images into line, word and characters. This is followed by feature extraction for representing character images and a classification module that label characters to their proper class. Finally, post processing i.e. applying algorithms such as spellchecker and other semantic approaches (Green, 1993).

There are four basic classification approaches in the pattern recognition literature. Since OCR is one of pattern recognition problem, these methods can be implemented in OCR. These methods are (Qing, 2003):Syntactic/Structural Pattern Classification, Statistical Pattern Classification and Mathematical Classification.Structural classification methods use structural features and decision rules to classify characters (Ralston et. al., 2000).Statistical pattern recognition relies on defining a set of decision rules based on standard statistical theory (Pandya and Macy, 1996).Many character recognizers are based on mathematical formalisms that minimize a measure of misclassification. These recognizers may use pixel-based features or structural features. Some examples are discriminant function classifiers, Bayesian classifiers, ANNs, and template matching. ANNs, which are closer to theories of human perception, employ mathematical minimization techniques. Both discriminant functions and ANNs are used in commercial OCR systems (Ralston et. al., 2000). Use of ANN systems, offer a new computing paradigm in which the network, through a process of learning from task examples can store experimental knowledge and make it available for use at a later time. These days, Europeans, Americans and others have been conducting researches and applying OCR technologies to their languages. As a result, these OCR technologies help to read different documents written in English, Chinese, Hindu, Arabic, Russian, and the like but do not read documents written in Amharic (Million, 2000).

This paper presents OCR engine for real life Amharic characters. The remaining part is organized as follows. Section 2 presents the characteristics of Amharic language with emphasis to its challenges to OCR development. The proposed system is discussed in Section3. Experimental results are presented in Section 4, and conclusion and future worksare highlighted in Section 5. References are provided at the end.

ISSN: 2278 – 7798

2. LINGUISTIC CHARACTERISTICS OF AMHARIC

2.1. AMHARIC LANGUAGE

Amharic, which belongs to the Semitic language, became a dominant language in Ethiopia back in history. It is the official and working language of Ethiopia and the most commonly learnt language next to English throughout the country (Million and Jawahar, 2005). The Amharic writing system was adopted from that of Geez. The Amharic writing system, by the time when it took over the place of Geez, took all the 26 Geez symbols and added several new ones to represent sounds not found in Geez. The additional symbols are □, □, □, □, □, □ and □. Since the 1st century, modifications and additions on Amharic characters has been undergone and current Amharic script has 33 basic characters, one special symbol in its 7 different forms to represent the |V| sound found in Latin based languages, 44 labial symbols, e.g. □,□,□, 8 punctuation marks and 20 symbols for numerals which make up a number of characters in Amharic writing to be greater than 330 (Bender et.al., 1976). In a syllabic system, like Amharic, the number of characters (symbols) needed by the language is determined by the number of basic sounds used (Bender et. al., 1976). In addition to the 231 characters there are nearly 44 others which contain a special feature usually representing labialization, for example, □ from □ and □ from □. Only about twenty of these are common and are usually listed as an appendix to the main list (Worku, 1997; Million, 2000).

2.2. FEATURES OF AMHARIC CHARACTERS

In a nutshell, the Amharic writing system has the following basic characteristics (Bender et. al., 1976; Nigussie, 2000):

- Each symbol is written according to the sound that goes with it. The vowels are integrated in the characters by modifying the base characters in some form, which together represent syllable combinations consisting of a consonant and vowel. Thus the Amharic writing system is often called syllabic rather that alphabetic.
- The symbols are written in a disconnected manner, e.g. □, □, □, □, □.
- Concerning the direction of writing the characters, while Sabaen characters are written right to left, Geez and Amharic and any language that uses the script for writing is written from left to right.
- The lines of text are read left-to-right and the lines are read in a top-to-bottom sequence.
- There is a proportional spacing between characters and the same is true for words.
- Words are delimited by wide white space either with in a line or at the end of lines.
- There is no capital, lower case distinction as it is for Latin characters.
- A line of Amharic printed script lies at the same level, having no ascent and descent features. This feature of Amharic writings system grants that there will always be a white line between two consecutive lines of characters, unless and otherwise the line is inclined.

2.3. CHALLENGES IN BUILDING AN OCR FOR AMHARIC DOCUMENTS

Character recognition from document images that are printed in Ethiopic script is a challenging task. To develop a successful character recognition system for Amharic script, the following issues need to be addressed.

2.3.1. DEGRADATION OF DOCUMENTS

Document images from printed documents, such as books, magazines and newspapers are extremely poor in quality. Popular artifacts in printed document images include (Million and Jawahar, 2005):

- Excessive dusty noise,
- Large ink-blobs joining disjoint characters or components,
- Vertical cuts due to folding of the paper,
- Cuts at arbitrary direction due to paper quality or foreign material,
- Degradation of printed text due to the poor quality of paper and ink,
- Floating ink from facing pages etc.

2.3.2. PRINTING VARIATIONS

Amharic printed documents vary in fonts, sizes and styles. Building character recognition system is challenging in this situation. For example, some of the commonly used fonts in Amharic printed documents include `PowerGeez', `VisualGeez', `Alphas', and `Agafari'. Each of these fonts offers several stylistic variants, such as normal, bold, italic, etc. They are also written in different point sizes, including 10, 12, 14, etc. These fonts, styles and sizes produce texts that greatly vary in their appearances within printed documents. Standardizing the variation in size by applying normalization techniques and extract suitable features is a challenging task so that the representation is invariant to printing variations.

ISSN: 2278 – 7798

2.3.3. LARGE NUMBER OF CHARACTERS IN THE SCRIPT

The total number of characters in Amharic script is more than 330. Existence of such a large number of Amharic characters in the writing system is a great challenge in the development of Amharic character recognizer. Memory and computational requirements are very intensive (Million and Jawahar, 2005). One needs to design a mechanism to compress the dimension of character representation so as to come up with computationally efficient recognizers.

2.3.4. VISUAL SIMILARITY OF MOST CHARACTERS IN THE SCRIPT

There are a number of very similar characters in Amharic script that are sometimes difficult for humans to distinguish them easily (examples are presented in Figure 1). Robust discriminant features needs to be extracted for classification of each of the character into their proper category or class.

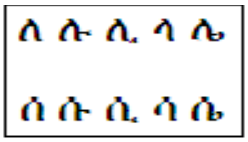


Figure 1.Samples of visually similar characters in Amharic writing system.

3. THE PROPOSED OCR SYSTEM

3.1. GENERAL ARCHITECTURE

The proposed Amharic OCR system is ANN approach.The system has three main components: preprocessing, segmentation, and recognition. Digitization, skew detection, noise removal, binarization, slant correction, thinning, normalization, and others belong to the preprocessing stage of the overall recognition system. In the segmentation step, the text image is separated into individual characters. Two essential components in a character recognition step are the feature extraction and the classification. The overall architecture of the system is shown in Figure 2.

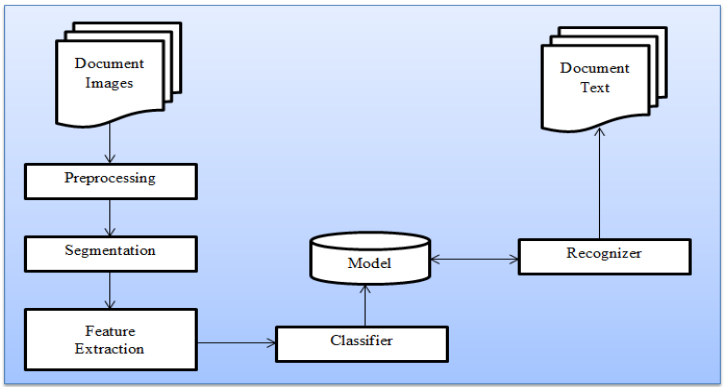


Figure 2.Architecture of the proposed OCR System

3.2. PREPROCESSING TECHNIQUES

In this component, noise removal and binarization techniques are applied. The adaptive filter is applied because it is more selective than a comparable linear filter, preserving edges and other high-frequency parts of an image.To binarize global image threshold, Otsu's method is used. This method is found to be more effective in isolating the text pixels from background pixels from the image without affecting the basic features of the character. The sample image before and after noise detection and removal techniques is shown in Figure 3 and Figure 4 respectively.

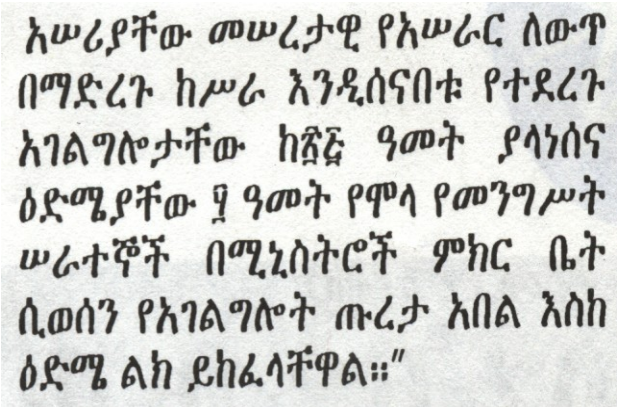


Figure 3. Sample Amharic text taken from "Federal NegaritGazeta" before noise removal techniques

ISSN: 2278 – 7798

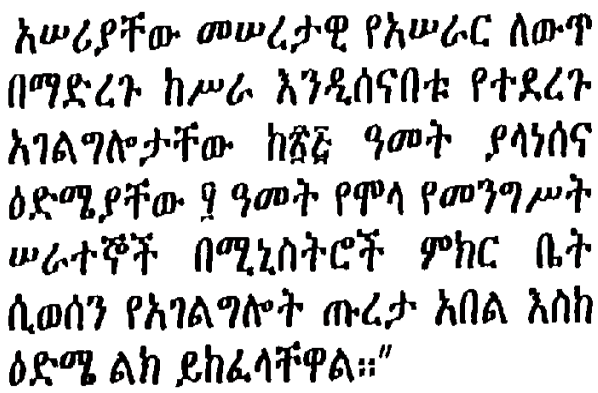


Figure 4. Sample Image after binarization and noise removal techniques

3.3. SEGMENTATION

The stage by stage line segmentation algorithm suggested by Pal and Chaudhuri (1995) is used and 100% accuracy is recorded. 98.55% of accuracy is recorded for character segmentation using the bounding box projection approach.

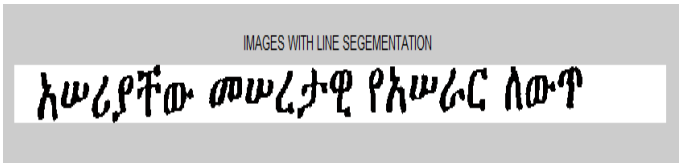


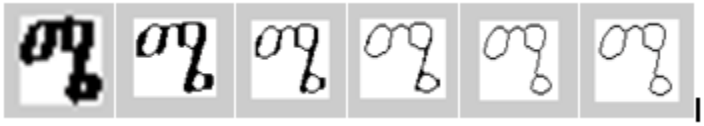
Figure 5. Sample Output of the Line Segmentation

(a)

(b)

Figure 6. Sample segmentation errors (a) the number "  " is taken as two distinct characters though it is not. (b) the four dots of "AratNetib" (□) and to strokes of quotation marks (") are recognized as distinct characters.

After character segmentation part is over, the normalizationand thinning preprocessing techniques are applied.Size normalization transforms each segmented character to some uniform height and width.A linear interpolation technique for normalization is applied. This data set is passed to thinning module in the next section and hit-and-miss morphological analysis for thinning are found to work very well for the problem of interest.



(a) (b) (c) (d) (e) (f)

Figure 7.Thinned characters with different iteration levels. (a) Original character and thinned at (b) 2, (c)3, (d)4, (e)5 and (f) Inf iteration levels.

3.4. RECOGNITION

ANN approach is used for character recognition. Neural networks are now increasingly used in printed character recognition applications (Haykin, 1994). ANN has four sub components: Design, Creation, Training and Testing.

Designing: The architecture of the neural network created for the recognition is a multilayer perceptron that accepts the binary representation of the segmented character as an input vector. The network created is a feed forward neural net with backpropagation algorithm. The approach adjusts the strength of connection between nodes at different layers after computing the error between the desired and the actual output of the network at each training iteration. The learning method is supervised learning where the desired output of the network for the know patterns of input is also fed to the network during training. By doing so the network takes inputs, performs some operation and provides the output using the feed forward approach. Then the mean squared error between the desired and the actual output is calculated. If this value is found to be greater than the threshold, the error propagated back to adjust the weights of each connection in the course of finding the optimal weight for each connation.

Creating: The network created is equipped with these two parameters, learning rate and momentum. The inclusion of learning rate is to specify how to make adjustment to the weights after calculating the errors on the output layer. Likewise, the purpose served by the momentum parameter is that, once the learning network knows to which direction to move on the learning curve, we can adjust the speed of the network for fast convergence.

Training: After creating the neural network, the next step is making the network to be introduced to different patterns of the whole character set. After different network parameters are

ISSN: 2278 – 7798

adjusted as per the performance of the network, the training input patterns along with matrix of target values are fed to the network for training.

Testing: In testing the network, two approaches were implemented. The first one is testing the network with the pattern that the network is trained with. This testing scenario helps to see if the network can recognize the original training pattern correctly. The second approach is evaluating the network performance with new input patterns. This testing scheme as well helps to see how powerful the network is in generalization when faced with new character patterns.

4. EXPERIMENT

4.1. DATA PREPARATION

The data collected are of two types. The first being data for training the neural network (recognition engine), the second is for testing the performance of the recognizer. Although, there are a number of sources to collect data from, such as Amharic books (fictions, sciences, arts, textbooks, etc); newspapers (private and government); and Amharic dictionaries, etc), it is considered reasonable to select representative real-life documents from different sources to increase the generalization ability of the ANN. Since the population size is unknown, sample size determination is a problem. The technique used for sampling is purposive sampling. It is decided to take a sample of four sources for the problem domain. These are texts from the popular government Amharic newspaper 'Addis Zemen', texts from the Amharic Bible, texts from the 'Federal NegaritGazeta', and the fiction 'FikerEskemekabir'.

The scanning dpi (density per inch) parameter for the sample documents is 300 dpi and with gray scale color format as recommended by Mori and Yamamoto, (1992). The scanned grayscale color images are then put as an input in the working folder to pass in to preprocessing techniques.After the normalized characters are thinned the binary representation of each character is created and saved to a file for recognition process.

The first step in training the network is rearranging the input pattern into a column vector containing elements equal to the number of neurons at the input layer. For this purpose, the binary representation of the characters extracted during the preprocessing stages, is converted into one column vector where each preceding column of the original character matrix will be appended to the previous one. For training the network, from eight training sets, a total of 1090 (1087 core and 3 labialized) characters are fed to the network. In testing, to create a match-up between the output and the target values, the 33 basic characters with their 6 forms (that is a total of 7x33 = 231) and 44 labialized characters, totally 275 characters, are assigned a number between 0 and 1 starting from the first character. The value of each fraction between 0 and 1 for 275 characters is assigned. Thus, for all the test cases what the network does is; it provides a fraction value between 0 and 1 as the number of output node is set 1 in the network definition. These fraction values are converted to its decimal equivalent value and the character with that decimal number representation is taken as an output of the network for that specific input pattern.To see whether the network can recognize the character set used for training, from eight training sets, a total of 1090 (1087 core and 3 labialized) characters are fed to the network. The network is also tested with new text for a total of 548 (543 core and 5 labialized characters) new characters, 542 of them are those included in the training set and the other 6 are found not included in the training set.

4.2. IMPLEMENTATION

For the purpose of implementing algorithms required to Amharic text recognition, MATLAB 7's Image Processing and Neural Network Toolboxes are used for the reason that it delivers increased performance and productivity by enabling developers to leverage existing skills. Besides, MATLAB's powerful collection of matrix manipulation routines which enables this paper to use matrixes, again closely used to model the real world objects (MathWorks, 2010).

4.3. TEST RESULTS

The errors encountered in the test results of the developed recognition system can be seen in two broad views: segmentation error and classification /recognition error. Segmentation error is an error occurred due to the segmentation algorithm implemented for this study. The segmentation error existed in the document was considering one character as more characters. This type of error exists in numbers and punctuations as shown in Figure 6.

The classification/recognition error is an error occurred when the developed system wrongly classifies/recognizes characters. The developed Amharic OCR system classified characters in to one of the following. The system:

- may recognize the characters correctly as expected.
- May recognize the characters incorrectly, but the output having similar shape with the expected.

ISSN: 2278 – 7798

- may recognize the characters incorrectly, with no similarity between the output and the expected.
- may classify the characters unknown (for characters not included in the training set).

In testing the developed system from the training set, from eight training sets, a total of 1090 (1087 core and 3 labialized) characters are fed to the network. The system correctly recognized averagely 96.87 % of the characters as shown in Table 1 below. That means, 3.13 % of the data was recognized incorrectly because of the existence of structurally similar characters. For instance, □ was recognized as □; □ was recognized as □.

On the other hand, in testing the developed system using new test set, on the average the system correctly classified 11.40% of the characters as shown in Table 2 below. Most of the characters are misclassified and basically two types of errors are observed. The first type of error happens because of shape similarity between characters. This error constitutes 73.6% of the total errors. For instance, □ was recognized as □. The second type of error is related to character deformation during highly corrupted image noise removal. For instance, X was recognized as x. This type of error accounts for 15% of the total errors.

Test Sets	Performance /Rate of Recognition
Training Set from Addis Zemen	97.88 %
Training Set from NegaritGazeta	97.84 %
Training Set from the Bible	96.41 %
Training Set from FikirEskemekabir	95.38 %
Average Accuracy	96.87 %

Table 1. Network Performance for Training Set

Test Sets	Performance /Rate of Recognition
Test Set from Addis Zemen	14.88 %
Test Set from NegaritGazeta	10.25 %
Test Set from the Bible	11.04 %
Test Set from FikirEskemekabir	9.44 %
Average Accuracy	11.40 %

Table 2: Network Performance for New Test Set

5. CONCLUSIONS AND FUTURE WORK

Character recognition from document images that are printed in Ethiopic script is a challenging task. To develop successful character recognition system for Amharic script, the degradation of documents, printing variations, large number of characters in the script and visual similarity of some characters need to be addressed. During experimentation of the applicability of preprocessing algorithms and approaches, the wiener adaptive filtering method for noise removal, Otsu global thresholding method for binarizing the digitized image, linear interpolation techniques for normalization and hit-and-miss morphological analysis for thinning are found to work very well for the problem of interest. In due course, the performance of the line segmenter is found to be 100%.The rate of segmentation for basic and labialized characters turns out to be 98.28% and 100% respectively for training character sets, 98.55% and 100% respectively for testing character sets.

For classifying the features generated, an ANN approach is implemented.The neural network is trained with eight samples taken from real-life documents. The performance of the developed system is tested with documents taken from real-life documents. Accordingly, an average recognition rate of 96.87% for the test sets from the training sets and 11.40% recognition rate is observed for the new test sets. The segmentation algorithm used in the current study worked reasonably for basic and labialized characters. But it fails to segment special character |v|, punctuations and numbers. In general, observation of the test results show that the performance of the system is greatly affected by the similarity of the shape of Amharic characters and effect of the application of noise removal for cleaning highly degraded document images. Such challenges require to further explore an invariant to shape feature extraction techniques and advanced noise detection and removal algorithms.

REFERENCES

1. Bender et *al*. (1976). ***"Language in Ethiopia"***: London, Oxford University Press,UK.
2. Genovese, J. A. (1970). ***"Character Recognition***": Encyclopedia of Library andInformation Science, Vol. 4, New York, Marcel DekerInc, USA.
3. Green, W. B. (1993). "***Introduction to Electronic Document Management System***": Boston, Academic Press Inc, USA.
4. Haykin S. (1994). ***"Neural Networks: A comprehensive Foundation"***: Prentice Hall, Inc., New Jersey, USA.
5. MathWorks, (2010). ***"MATLAB 7 Image Processing Users' Guide"***: The MathWorks, Inc,USA.
6. Million Meshesha (2000). "***A Generalized Approach to Optical Character Recognition of Amharic Texts***": (Masters Thesis), School of Information studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia.
7. Million Meshesha and Jawahar, C. V. (2005). "***Recognition of Printed Amharic Documents***": Proceedings of Eighth International Conference on Document Analysis and Recognition (ICDAR), Vo. 21, No.8-10, pp. 784-788.

ISSN: 2278 – 7798

8. Mori, S., Suen, C., and Yamamoto, K. (1992). "***Historical Review of OCR Research and Development***": Proceedings of the IEEE, Vol. 80 No.7, pp. 1029-1058.
9. NegussieTaddesse (2000). "***Handwritten Amharic Text Recognition Applied to the Processing of Bank Cheques***": (Masters Thesis), School of Information Studiesfor Africa, Addis Ababa University, Addis Ababa, Ethiopia.
10. Pal and Chaudhuri (1995). "***A Complete Printed Bangla OCR System"***: Pattern Recognition, Vol.31, No.5, pp.531-549.
11. PandyAbhijits S., Macy Roberts B. (1996). ***"Pattern Recognition with Neural Networks in C++"***: CRC Press,USA.
12. Qing Chen (2003). ***"Evaluation of OCR Algorithms for Images with Different Spatial Resolutions and Noises"***: Journal of the Research Institute for Computer and Information Communication, Vol.15, No.3, pp. 45-51.
13. Ralston *et al.* (2000). ***"Optical Character Recognition"***:Encyclopedia of Computer Science, 4th ed., Nature Publishing Group,USA.
14. WorkuAlemu (1997). "***The Application of OCR Techniques to the Amharic Script***":(Masters Thesis), School of Information Studies for Africa, Addis AbabaUniversity, Addis Ababa, Ethiopia.

ISSN: 2278 – 7798