Ethiopic Document Image Database for Testing Character Recognition Systems

Yaregal Assabie and Josef Bigun

School of Information Science, Computer and Electrical Engineering
Halmstad University, SE-301 18 Halmstad, Sweden
{Yaregal.Assabie, Josef.Bigun}@ide.hh.se

Abstract. In this paper we describe the acquisition and content of a large database of Ethiopic documents for testing and evaluating character recognition systems. The Ethiopic Document Image Database (EDIDB) contains documents written in Amharic and Geez languages. The database was built from a variety of documents such as printouts, books, newspapers, and magazines. Documents written in various font types, sizes and styles were included in the database. Degraded and poor quality documents were also included in the database to represent the real life situation. A total of 1,204 pages were scanned at a resolution of 300 dpi and saved as grayscale images of JPEG format. We also describe an evaluation protocol for standardizing the comparison of recognition systems and their results. The database is made available to the research community through http://www.hh.se/staff/josef/.

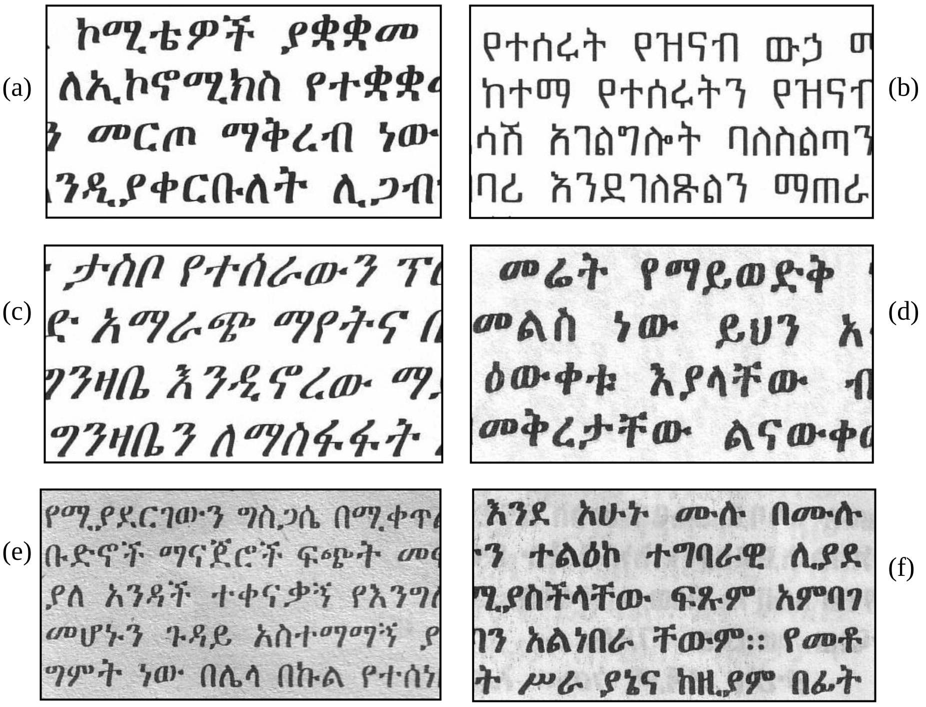
1. Introduction

With the growing need of computerized information processing, many researches in the area of recognition of various scripts have been conducted over the past decades. The layout and qualities of documents used for testing recognition systems highly affects the results. Thus, to standardize and compare research results, document image databases are built for various scripts. Examples include NIST for Latin [1], CEDAR for Japanese [2], ERIM for Arabic [3], etc.

However, there has no standard database for Ethiopic text so far. As part of the research on recognition of Ethiopic characters, we developed Ethiopic Document Image Database (EDIDB). The database is also intended to serve other researchers in the area and standardize the research on Ethiopic character recognition. EDIDB helps researchers to test recognition systems with respect to variation in font type, font style, font size, skewness, text uniformity and document type.

2. EDIDB Specification

Real life Ethiopic documents written in Amharic and Geez languages are collected for database development. The database consists of grayscale images of 1,204 pages saved in JPEG format. All documents are scanned with CanoScan LiDE 20 flatbed scanner at a resolution of 300 dpi. Documents are classified into four major groups: *printouts*, *books*, *newspapers* and *magazines*. The qualities of images vary from high quality printout documents to low quality books, newspapers, and magazines which are published as far as before 15 years. Sample documents are shown below in Fig.1.



ኮሚቴዎች ያቋቋመ
ለኢኮኖሚክስ የተቋቋ
ን መርጠ ማቅረብ ነው
ንዲያቀርቡለት ሊጋብ

የተሰሩት የዝናብ ውኃ
ከተማ የተሰሩትን የዝናብ
ሳሽ አገልግሎት ባለስልጣን
ባረ እንደገለጹልን ማጠራ

ታስቦ የተሰራውን ፕ
ድ አማራጭ ማየትና በ
ግንዛቤ እንዲኖረው ማ
ግንዛቤን ለማስፋፋት

መሬት የማይወድቅ
መልስ ነው ይህን አ
ዕውቀቱ እያላቸው ብ
መቅረታቸው ልናውቀ

የሚያደርገውን ግስጋሴ በሚቀጥ
ቡድኖች ማናጀሮች ፍጭት መ
ያለ አንዳች ተቀናቃኝ የእንግ
መሆኑን ጉዳይ አስተማማኝ ያ
ግምት ነው በሌላ በኩል የተሰ

እንደ አሁኑ ሙሉ በሙሉ
ን ተልዕኮ ተግባራዊ ሊያደ
ሚያስችላቸው ፍጹም አምባገ
ነን አልነበራ ቸውም:: የመቶ
ት ሥራ ያኔና ከዚያም በፊት

Figure 1. Sample documents from EDIDB showing (a) VG2000 Main font, (b) VG2000 Agazian font, (c) Visual Geez Unicode font with italic style, (d) book, (e) newspaper, (f) magazine

A summary of the specific details of the scanned documents are presented below.

2.1. Printouts

- Total scanned pages: 983
- Fonts Types: Geez Type, Power Geez, Visual Geez 2000 Main, Visual Geez 2000 Agazian, and Visual Geez Unicode
- Font sizes: 8, 10, 12, 16, 20
- Font styles: Normal, Italic, and Bold
- Skewness: Non-skewed, and skewed from -30o to 30o
- Uniformity of text: Uniform in font size and style, and combination of various font sizes (8 to 20) and styles (normal, bold, italic, and bold+italic)

2.2. Books

- Total number of books: 5
- Total scanned pages: 116

2.3. Newspapers

- Total number of newspapers: 3
- Total scanned pages: 79

2.4. Magazines

- Total number of magazines: 2
- Total scanned pages: 26

3. Evaluation Protocol

One objective of using the same database for testing character recognition systems is to be able to compare the results. This is achieved by setting a standard evaluation protocol in which researchers are expected to follow as a guideline for reporting and evaluating their results. Therefore, we describe the following evaluation protocols for test results reported by using EDIDB.

3.1. Font Types

EDIDB offers researchers who want to test their recognition systems with variations in font types. A text of 83 pages in *Visual Geez Unicode* font was prepared in other four font types (*Geez Type*, *Power Geez*, *Visual Geez 2000 Main*, and *Visual Geez 2000 Agazian*) with the same font size of 12. Recognition results can be reported and comparisons can be made between different font types as follows.

Font type	Geez Type	Power Geez	Visual Geez 2000 Main	Visual Geez 2000 Agazian	Visual Geez Unicode
Recognition					

3.2. Font Sizes

EDIDB can also be used to evaluate recognition systems under variations in font size. A copy of the *Visual Geez Unicode* document (83 pages with 12 font size) was prepared with font sizes of 8, 10, 16 and 20. Recognition results can be reported and comparisons can be made between different font sizes as follows.

Font size	8	10	12	16	20
Recognition					

3.3. Font Styles

EDIDB has document images with three font styles: normal, italic, and bold. A copy of the *Visual Geez Unicode* document (of 83 pages in normal style and 12 font size) was prepared with italic and bold styles. The results of recognition systems can be compared across different font styles as follows.

Font style	Normal	Italic	Bold
Recognition			

3.4. Skewness

EDIDB also offers test data for researchers who want to test their recognition systems with skewed documents. The *Visual Geez*

Unicode document (of 83 pages in normal style and 12 font size) was artificially skewed to various inclination angles of the interval [-30^o,30^o]. The inclination angles for each image are encoded in their last three alphanumeric strings of image file names, where the first character indicates direction (N for negative and P for positive) and the next two are digits showing the angle. For example, an image whose file name is "Unicode12Skewed-0016P30.jpg" has an inclination of 30^o, and an image whose file name is "Unicode12Skewed-0005N14.jpg" has an inclination of -14^o. The tolerance of recognition systems to skewness can be reported by describing the inclination angle along with the results. Comparison with non-skewed documents can be made by testing the recognition system with the Visual Geez Unicode document with 12 font size and normal style. The following table can be used as a general format for reporting results.

Skewness	
Recognition	

3.5. Uniformity of Text

In real life documents, the format of a text in a page may not be uniformly formatted. To represent this reality, a copy of the *Visual Geez Unicode* document (of 83 pages in 12 font size and normal style) was prepared by changing the formats. This non-uniform document contains mixed font sizes of 8 to 20 and styles with normal, bold, italic, and italic+bold. Then, the tolerance of a recognition system on non-uniform documents can be compared with the uniform *Visual Geez Unicode* document and the results can be reported in the following format.

Uniformity	Uniform	Non-uniform
Recognition		

3.6. Document Types

The results obtained by recognition systems usually vary due to variations in the document types. The quality of documents has a direct effect on the results. To compare and contrast the results on various document types, a test can be made on EDIDB images taken from printouts, books, newspapers and magazines. These images vary from degraded and poor quality images to good quality printouts. It is possible to compare results within each document type or the following general format can be used for reporting (and comparing) the average results of each document type.

Document type	printouts	books	newspapers	magazines
Recognition				

References

[1] http://www.nist.gov/srd/optical.htm
[2] http://www.cedar.buffalo.edu/Databases/JOCR/
[3] http://documents.cfar.umd.edu/resources/database/ERIM_Arabic_DB.html