1. BACKGROUND

Optical Character Recognition (OCR) is a technology that is used to translate scanned images of text into computer editable and searchable text.

Among others, the following are the major advantages of the OCR technology:

- It can be used to scan and preserve historical documents.
- It can be used for scanning data entry forms in a faster and less error prone manner.
- It can be used with other computer applications, such as Archives and Records Management Systems, to convert scanned documents into searchable text.

At present the recognition of Latin-based characters from well-conditioned documents can be considered as a relatively feasible technology. On the other hand, the processing of non-Latin scripts is still a subject of active research.

Ethiopic script-based OCR processing is currently among the least developed ICT disciplines in the country. Developments in this area are mainly limited to preliminary research activities undertaken at different institutions of higher educations, such as the former School of Information Science for Africa (SISA). Such efforts are undertaken in an uncoordinated ways.
2. OCR RESEARCHES IN ETHIOPIA

Our team has tried to review the local efforts undertaken in the country in this area. At present no single production quality or commercial OCR application exists that processes Ethiopic documents.

Some students of the departments of Information Science, Computer Science and Electrical and Computer Engineering of the Addis Ababa University have produced research output that is of interest to the fields of Optical Character Recognition Software development to Ethiopic documents.

The students have used different algorithms to reach to a better result in the recognition of characters from scanned documents. Some of the algorithms and techniques that we reviewed are:

- Character Recognition based on polygonal approximation (mathematical) and topological features (The Application of OCR Techniques to the Amharic Scripts: By Worku Alemu)
- Thinning algorithm (A generalized Approach to Optical Character Recognition (OCR) of Amharic Texts: By Million Meshesha)
- Training with Neural Networks (Development of a versatile Character recognition system for Amharic texts: By Yaregal Assabie)

The researches have encompassed both typewritten and computer written document types. The studies show that those tests done on the latter type of documents produce a better result.

It is the recommendation of all researchers in this area that, to achieve a better result in OCR as applied to the Ethiopic script, a national standard font and keyboard mappings must
be designed and developed. This will help the test results that are obtained through the learning process (e.g. Neural Networks) to produce better results as the variation of font types and styles becomes less.

Although the efforts of such researches which are undertaken as a partial fulfillment to the requirements of the various degree programmes are commendable, there is a clear lack of coordinated effort that would take such researches beyond academic exercises.
3. INTERNATIONAL MULTILINGUAL OCR TOOLS VENDORS

One of the steps that the Government may consider taking in order to develop the OCR technology is to work in partnership with international vendors that develop multilingual OCR processing tools.

We have identified some of the major developers of OCR Software Development Kits (SDK) in order to identify the requirements of incorporating additional scripts in such SDKs.

One of the major inputs demanded by such vendors is the availability of a standard Typeface that would enable the OCR tools learn the features of the Ethiopic characters. The output of this project can be considered as step forward in this regard.

The following are among the major providers of OCR Software Development Kits.

1. Microsoft Office Document Imaging Library (MODI)
2. OmniPage Capture SDK
3. Leadtools OCR Programming Tools

The focus in this regard on the identification of Programmable tools that provide interfaces for customization, as opposed to end-user OCR products.

3.1. Microsoft Office Document Imaging Library (MODI)

Microsoft has recently incorporated an OCR processing technology in its MS Office 2003 packages known as Microsoft Office Document Imaging Library (MODI).
MODI provides OCR Application Development Interfaces (API) that allows developers incorporate OCR capabilities into their products.

Although MODI does not provide support for Ethiopic, it provides support to other non-Latin scripts. The present version of MODI supports the following languages:

- Chinese
- Czech
- Japanese
- Korean
- Russian, and
- A number of European Languages

One of the advantages of MODI is that it is freely available to developers that have installed an MS Office product.

3.2. **OmniPage Capture SDK**

OmniPage Capture SDK is a popular product that supports OCR processing capabilities to non-European languages such as the Japanese, Chinese and Korean scripts.

The OmniPage Capture SDK provides different OCR engines such as print OCR (OCR, OCR-A, OCR-B and MICR), Handprint (ICR), Check Mark (OMR) and Barcode recognition engines. It also provides image file enhancement tools as well as facilities for exporting processed outputs to different formats such as PDF and XML.
3.3. **Leadtools OCR Programming Tools**

Leadtools OCR Programming Tools is another popular OCR SDK that supports multilingual OCR processing.

It provides facilities for exporting processed output into different file formats; provides different OCR engines; and includes image file enchantment facilities.
4. **COMMON STEPS OF OCR PROCESSING**

The process of converting documents into electronic forms, which is usually referred to as digitization is undertaken in different steps.

The process of scanning a document and representing the scanned image for further processing is called the pre-processing or imaging phase.

The process of manipulating the scanned image of a document to produce a searchable text is called the OCR processing stage.

4.1. **The Imaging Stage**

The imaging process involves scanning the document and storing it as an image. The most popular image format used for this purpose is called Tagged-Image File Format (TIFF).

The resolution (number of dots per inch - dpi) determines the accuracy rate of the OCR process.

4.2. **The OCR Process**

The major steps of the OCR processing stage are shown below.

4.3. **Distinguishing between text and images - Segmentation**

In this step, the process of identifying the text and image blocks of the scanned image is undertaken. The boundaries of each image are analyzed in order to recognize the text.

4.4. **Character recognition - Feature Extraction**

This step involves recognizing a character using a method known as feature extraction. OCR tools store rules about the characters of a given script using a method known as the learning process. A character is then identified by analyzing its shape and
comparing its features against a set of rules stored on the OCR engine that distinguishes each character.

4.5. Recognition of Words

Following the character recognition process, word identification process is performed by comparing the string of characters against an existing dictionary of words. Additional processes such as spell-checking are performed under this step.

4.6. Correction of Unrecognized Characters - Error Correction

In this step, the user is allowed to provide corrections to unrecognized characters.

4.7. Output Formatting

The final step involves storing the output in one of the industry standard formats such as RTF, PDF, WORD and plain UNICODE text.
5. GUIDELINES FOR SELECTING AN OCR PROCESSING PRODUCT

The following are widely accepted guidelines for selecting high quality OCR processing products. The guidelines can be particularly useful if the Government considers the possibility of working with international OCR Software Development Kit vendors.

The guidelines should be applied under similar conditions in terms of the quality and size of documents to be scanned, the quality of the scanning hardware and the resolution under which the document was scanned.

5.1. The Number and Type of OCR Engines Available

Modern OCR software uses multiple engines to achieve a high level of accuracy. Such engines include Handprint (ICR), machine print OCR (OCR-A, OCR-B, etc), Check Mark (OMR) and Barcode (1D and 2D) recognition engines.

5.2. Recognition Speed

The speed at which the OCR software recognizes a given scanned document is an important OCR software selection factor.

5.3. Image Testing Process

This criterion refers to the ability of the OCR software in converting printed documents to computer application files such as RTF, MS WORD, and PDF. For example, the OCR output may be saved as rtf file (Rich Text Format document) and compared to the original physical document. The number of errors in the scanned document is then calculated for each OCR software.

5.4. Supported Output Formats

The number of possible output formats such as XML, HTML, PDF, and DOC in which the output of the
scanned documents can be saved is also an important selection factor.

5.5. Support for Unicode Fonts
The OCR software under consideration should also support Unicode fonts that consist of glyphs of different international character sets including Ethiopic.

5.6. File Enhancement Features
The quality and conditions of the original documents affects the OCR processing process. The OCR software must have facilities for removing discolorations and improving contrast. Such paper color removal and contrast adjustment enhancement should not affect the accuracy level of the OCR processing.

5.7. Availability of advanced features
The availability of advanced features such as spell-checkers and WYSIWYG editors should also be considered as one of the selection criteria.
Although, there exist a number of methodologies of addressing the issues of character recognition we have found that the Neural Networks approach is nowadays the most popular method for recognizing non-Latin scripts such as Ethiopic. This is a research area that the EICTDA may consider working with researchers from the various academic institutions as well as other private bodies.

Neural Network is one of the areas of Artificial Intelligence that has gained popularity in multilingual Optical Character Recognition. In particular, the Backpropagation neural network is one of the popular approaches employed for this problem area.

A neural network consists of three interconnected layers - Input, Hidden and Output layers. The processing nodes of each layer are tied together with weighted links.

The main approach is that a training set is first prepared, and then a neural network is trained to recognize patterns from the training set. The training process teaches the network to respond with desired output for a specified input. Each training sample is represented by possible input and the desired network's output for the input. Finally, the network is expected to produce the desired output for a given arbitrary input.

To apply the process to character detection, the training process involves storing a large number of character images with their corresponding Unicode character. The training set is used during the character recognition stage of the OCR process.

The above discussion of the Neural Networks approach provides only a brief outline of the overall methodology. We recommend that the EICTDA support through researches in this area.
7. RECOMMENDATIONS

Although the OCR technology is one of the important areas of the Computer technology, our assessment reveals that no, or very minimal, effort is being undertaken in this area. Thus the Ethiopian Government should put efforts towards the development and promotion of the OCR technology.

Apart from the very few graduation thesis outputs, there exists on production quality Ethiopic OCR application.

To promote the development of the Ethiopic OCR technology, the following recommendations are provided.

7.1. **Enforce the use of national standard typeface.**

The use of a standard typeface across organizations would make the processing of Ethiopic documents much easier, as the attributes of such typeface would be pre-known to OCR processing applications. Thus Government has to undertake promotional activities regarding the designed standard typeface and eventually enforce the use of the standard typeface for the exchange of documents among Governmental organizations.

7.2. **Support OCR Research Activities.**

In order to achieve a sustainable outcome, the Government has to consider this area as one of its priority research components. A coordinating body should be setup to oversee the research outcomes undertaken by academic institutions, private bodies as well as Governmental intuitions. A public resource center should also
be setup to promote and disseminate the outcomes of the research activities.

7.3. **Outsource International OCR Software Development Kits.**

At present there exist a number of international OCR Software Development Kit (SDK) vendors that support the processing of Multilingual Documents, including documents created using Japanese, Chinese and Korean scripts.

The EICTDA needs to work in partnership with the International vendors in order to incorporate the Ethiopic script in such products. The availability of a standard typeface and keyboard layout and the fact that the Ethiopic script is part of the Unicode standard character set is expected to make the inclusion of Ethiopic OCR processing in the popular products far easier.

The selection criteria presented in section 5 should be used to select the most appropriate product.
Contents

1. Background................................................................. 1
2. OCR Researches in Ethiopia........................................... 2
3. International Multilingual OCR Tools Vendors ............ 4
   3.1. Microsoft Office Document Imaging Library (MODI) .. 4
   3.2. OmniPage Capture SDK........................................... 5
   3.3. Leadtools OCR Programming Tools......................... 6
4. Common Steps of OCR Processing ................................. 7
5. Guidelines for Selecting an OCR Processing Product ... 9
   5.1. The Number and Type of OCR Engines Available... 9
   5.2. Recognition Speed ............................................... 9
   5.3. Image Testing Process......................................... 9
   5.4. Supported Output Formats .................................... 9
   5.5. Support for Unicode Fonts ................................. 10
   5.6. File Enhancement Features ................................. 10
   5.7. Availability of advanced features ...................... 10
6. Character Recognition using Neural Networks........... 11
7. Recommendations ......................................................... 12