DESIGN AND DEVELOPMENT OF AN OCR SYSTEM THAT CAN CONVERT BOTH AMHARIC AND ENGLISH BASED IMAGES AND SCANNED PDF FILES INTO EDITABLE CONTENT

Tibebe Beshah, Addis Ababa University
Abraham Asfawosen, Addis Ababa University

ABSTRACT

Optical Character Recognition (OCR) is technology of recognizing printed or written text characters by a computer. It is being used in many areas like libraries and information centers to document and preserve handwritten and/or computer processed text images. Done effectively it is assumed to facilitate text recognition and processing. Though there are attempts in building such systems, their accuracy and applicability due to language difference is limited. Thus, through this research attempt is made to investigate and develop an OCR system that can recognize both Amharic (local language of Ethiopia) and English text images. Design science research process is followed through out the research and a convincing result is achieved.

Keywords: Optical Character Recognition (OCR); Amharic; Design Science; Methods.

INTRODUCTION

OCR (optical character recognition) is the recognition of printed or written text characters by a computer. This involves photo scanning of the text character-by-character, analysis of the scanned-in image, and then translation of the character image into character codes, such as ASCII, commonly used in data processing. In OCR processing, the scanned-in image or bitmap is analyzed for light and dark areas in order to identify each alphabetic letter or numeric digit. When a character is recognized, it is converted into an ASCII code. Special circuit boards and computer chips designed expressly for OCR are used to speed up the recognition process.

OCR is being used by libraries to digitize and preserve their holdings. OCR is also used to process checks and credit card slips and sort the mail. Billions of magazines and letters are sorted every day by OCR machines, considerably speeding up mail delivery.

It is Widely used as a form of information entry from printed paper data records – whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation – it is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

Currently most organizations have a lot of written resources like books or magazines. When those organizations loose those resources from their local pc but they need that particular resource they are enforced to type from the scratch since they have no any mechanism to get the

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

original softcopy for modifying the resource. But using OCR system this can be handled easily since they only need to take the photograph of the resource and feed the image into this system to get the actual text they see on the image.

STATEMENT OF THE PROBLEM

The major reason for developing OCR application is to solve the following problems. The first one is that currently most of OCR applications developed do not work for Amahric language rather most of them work for other foreign languages like English. The other problem is that most organizations have a lot of written hard copy materials which are not found in soft copy so the only way they have when they need to modify some content and use the resource is to type the whole content from scratch for minor modifications which takes a lot of time and effort. So the application is developed to solve those problems in easy and efficient manner to solve problems faced by organizations while using previous written resources.

In relation to this attempt was made to explore the possibilities of developing an OCR system for typewritten Amharic text by Dereje (1999). In this research previous algorithm implemented for recognition of Amharic characters is modified to incorporate the specific features of typewritten Amharic characters. The segmentation and the noise removal algorithms are integrated with this algorithm. It is then tested on five pages of typewritten Amharic text with isolated characters and 61% accuracy is registered. Further research attempts are recommended to improve the performance.

Another research by Meshesha & Jawahar (2000) proposed a novel feature extraction scheme using principal component and linear discriminant analysis, followed by a decision directed acyclic graph-based support vector machine classifier. In this research on an average around 90 percent recognition is obtained as reported by the authors. Though it is relatively better compared to other works in the area it still needs further improvement.

There are also other attempts on foreign languages. Basich & Sturzenegger (2017) conducted research with key objective of developing a character recognition system on images for Japanese language combining standard image segmentation and classification technique. The research aims at developing a system that is capable of recognizing both typeset and hand written characters of any font type with better result.

Authors used Python using Open CV as its image processing library, SciPy and NumPy for sophisticated mathematical processing, and the scikit-learn package for machine learning. By leveraging a user-friendly language that easily calls into C and C++ functions, our solution is reasonably efficient while also allowing for rapid development. The basic techniques used for developing the system include Segmentation and Classification. Segmentation is the heuristic used in image segmentation assumes that lines of text are all oriented in the same direction. Furthermore, it relies on the text being fixed-pitch, meaning that each character is given close to the same space regardless of the size of the character itself. Segmentation is performed on a binary version of the image: pixel values over the threshold are black and those under the threshold are white. The segmentation process occurs in three steps. First, the orientation of the image must be normalized so the kanji are not rotated and each line of characters is strictly horizontal. The image is then cut into lines delimited by whitespace in the vertical direction. Finally, each line is processed so each character is segmented so it can be classified with the minimum amount of noise. Classification is the process where the candidate image is compared against historically observed golden image-kanji pairs in order to select a most likely kanji for

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

the image in question. Small candidate kanji images are cropped out of the original documents image by the segmentation engine and flow into the classifier. The classifier has two distinct stages: first the training step and then the classification stage. The system uses a classic machine learning method: Gaussian Naive Bayes.

According to the authors key outcome of the research includes competitive with the free, online tool used for baseline reports for some fonts. By leveraging a large feature space, the system is able to correctly extract Japanese characters from standard documents. Further research opened opportunities to improve the system in the fields of handwriting recognition and training classifiers on larger corpora of text (Das & Banerjee, 2012; 2014).

Al-A'Ali & Ahmad (2007) conducted research to develop a new technique based on feature extraction and on dynamic cursor sizing for the recognition of Arabic Text. The research aims at solving the most challenging area in Arabic OCR (AOCR) research i.e. the segmentation of words into their sub-words and their individual characters by applying rules that govern the size and movement of the cursor through each segment. The research aims at solving problems that exist in the training of Arabic fonts especially at the segmentation stage.

The basic techniques the researcher applied in order to develop the OCR model includes segmentation, Feature Extraction, Definition and Representation and character classification. The key outcome of the research, as reported by the authors, is a new approach for Arabic character recognition based on producing a logical dynamically sized cursor to traverse the image of the Arabic word (Zayene et al., 2018). The research demonstrates that the identifying the directional vectors of the strokes of the Arabic characters within the word is the best way forward for Arabic OCR's. The research opens way to investigate more on how to recombine strokes into characters where each stroke should be added to neighboring strokes through available connection points until it fits into a character class by building a database of character classes such as characters that have a closed loop or a half open upwards circle is another future research direction. Thus, it can be seen that there are limited researches on Ethiopian languages and context though there are attempts in all languages.

Accordingly, the general objective of the research is to develop OCR application that translates both Amharic and English based images and scanned pdf files into editable content. More specifically the system is expected to perform translation of Amharic based images into normal text, perform translation of English based images into normal text, perform translation of Amharic and English based images into editable text, and perform translation of scanned pdf documents into editable text.

METHODOLOGY

Generally, a design science research approach based on Offermann et al. (2009) and Chaudhuri et al. (2020) Wu & Wang (2020) process model is applied. Accordingly, the research process followed three major stages as described below. They are Problem identification, Solution Design and Evaluation.

Problem Identification: In this phase of the research process, we identified and justified that the problem worth considering in the sense that there is scarcity of research on Ethiopian languages written text image recognitions through literature review and preliminary observation. The relevance has been pointed out in various literatures that we reviewed (Figure 1).

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

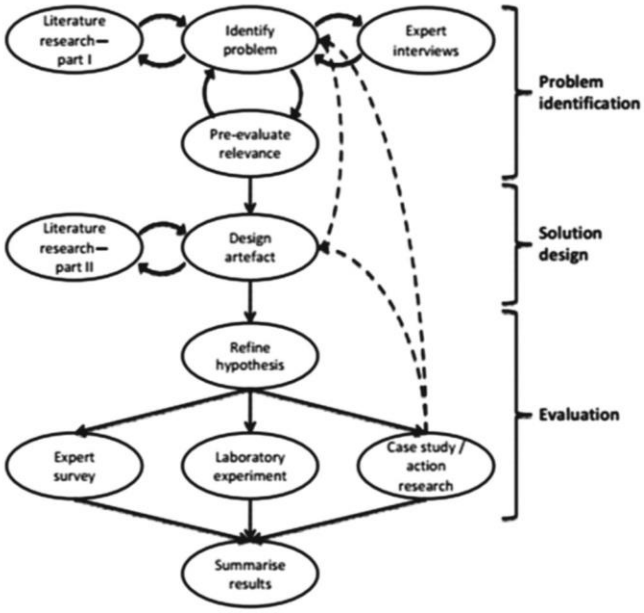

FIGURE 1
RESEARCH PROCESS

(Source: Philipp Offermann et al., 2019)

Solution Design

At this stage basic process like requirement/feature determination, model and experimental designs as well as development of the actual solution has been done. Accordingly, UML models like use case and activity modeling are used. Relevant literatures were also consulted in the process. During the development various component of the artefact are iteratively and meticulously experimented. In order to develop the application, the following tools and technologies are used:

NetBeans

NetBeans is an Integrated Development Environment (IDE) for Java. NetBeans allows applications to be developed from a set of modular software components called modules. The IDE simplifies the development of web, enterprise, desktop, and mobile applications that use the Java and HTML 5 platforms.

Tesseract

Tesseract is an open source optical character recognition (OCR) platform. OCR extracts text from images and documents without a text layer and outputs the document into a new searchable text file, PDF, or most other popular formats. Tesseract is highly customizable and can operate using most languages, including multilingual documents and vertical text (Sai Abhishek et al., 2021). Although the software can be used on Windows or Linux, this guide will be based on Mac operating systems which are done through the terminal application.

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

jTessBoxEditor

It is a box editor and trainer for Tesseract OCR, providing editing of box data of both Tesseract 2.0x and 3.0x formats and full automation of Tesseract training. It can read images of common image formats, including multi-page TIFF. It is also helpful to improve verification code recognition rate, tesseract training samples.

Evaluation

At this stage, our evaluation has been made using expert evaluations. We roll out the system and measure the applicability and use of the artifact and complied the results. Moreover a comparison was made with the relevant related work to ensure its rigor.

PROPOSED SOLUTION AND RESULTS

Description of the Envisaged System

Based on the research problem, it is identified that the envisaged OCR system has to have key features like performing the conversion of images into text through appropriate logic to perform its task and accurate validations for the application. In doing the main task the system is also expected to exhibit features like attractive user interface, incorporate dpi optimizer for assigning dpi value for images, being easy to use and maintainable.

The following Figure 2 shows the ahigh level representation of the system working as proposed in the OCR system.

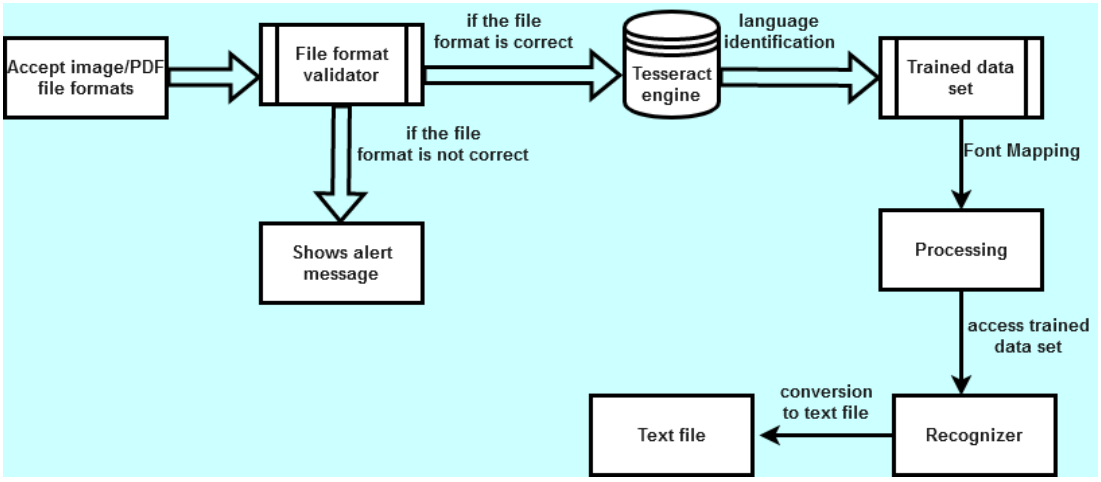

FIGURE 2
HIGH LEVEL ARCHITECTURE OF THE SYSTEM

System Models

Use Case Diagram along with a description, activity diagram and sequence model of the envisaged system is presented below. Accordingly Table 1 presents functionality of the proposed

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

system and it is followed by a description of the system through use case documentation table (Table 1).

Table 1
USE CASE DOCUMENTATION

Use Case Id	OCR-001
Use Case name	Object character recognition
Use Case description	The usecase describes the processes performed by the system to perform translation of images into text from accepting of images from the user up to converting the image into text.
Actors	Actors of the system are: User: anyone who enter the system to convert image into text.
Pre-conditions	The user is expected to select language and image file.
Flow of events	1. The user selects language of the image. 2. The user selects the image or pdf file. 3. The system will convert the image/pdf into text.
Post conditions	The converted text will be saved as a text file on the users' pc.
Alternate flow	3.1 The system checks weather the user select language or not. 1. If so the system will allow the user to choose file 2. Else the system will disable the file chooser. 3.2 The system checks weather the user select image/pdf file or not. 1. If so the system performs the conversion process. 2. Else the system will display error notification message.

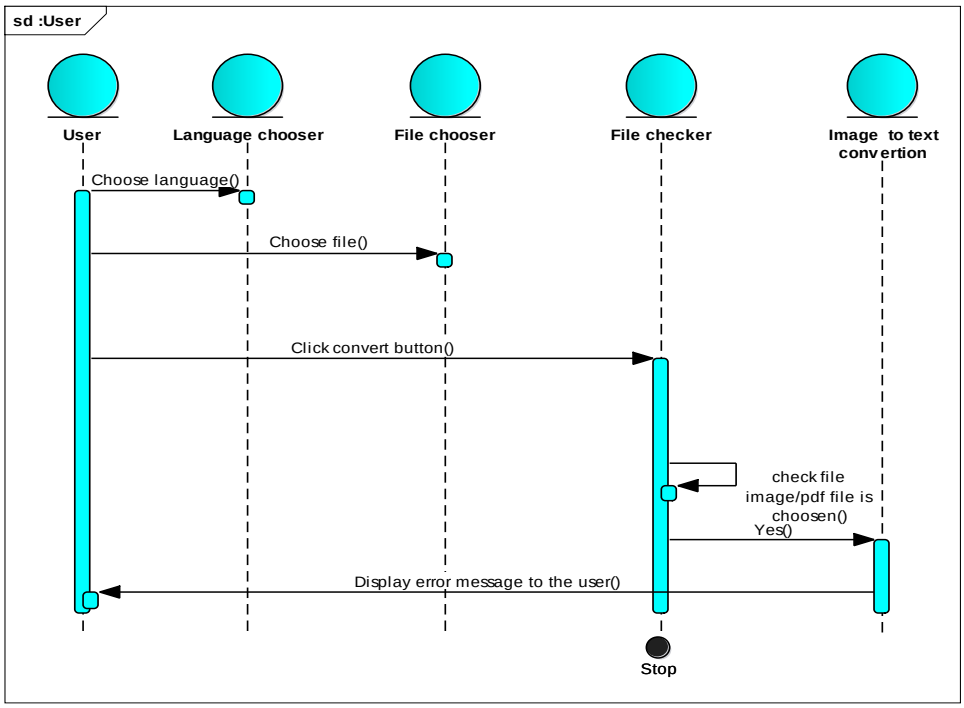

FIGURE 3
SEQUENCE DIAGRAM FOR OCR

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

Sequence diagram (Figure 3) representation for the OCR system for better presentation of how it works is shown below.

Activity diagram (Figure 4) representation for the OCR system for better presentation of how it works is shown below.

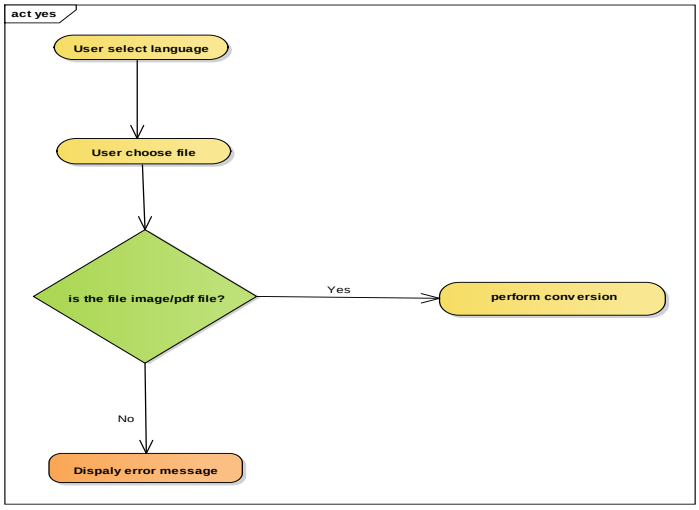

FIGURE 4
ACTIVITY DIAGRAM FOR OCR

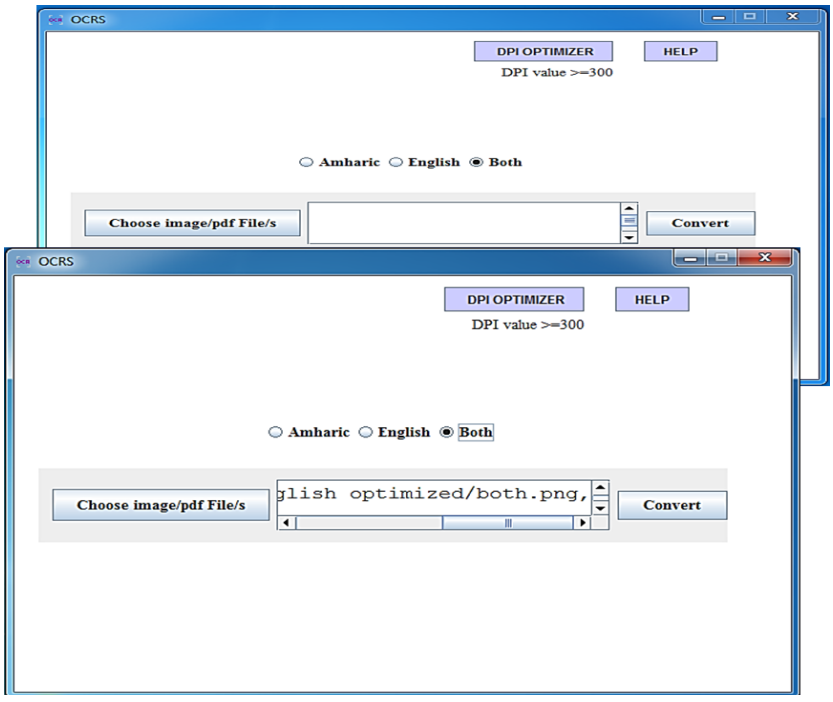

FIGURE 5
SAMPLE INTERFACE

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

Solution Development

The implementation of the proposed OCR system is presented in terms of its interfaces and sample codes. Accordingly, sample interfaces (Figure 5) are presented below.

Some key tasks of the OCR system are presented with sample code here below. Accordingly, before starting to convert image files to editable text for both Amharic and English based images the system first checks if the selected file is image file format or not. For this case let us assume the image is Amharic based. Then the system will create a folder c/ocr/Amaharic automatically to store the result to text file as shown below (Figure 6).

```
if(h[i].contains(".JFIF")||h[i].contains(".jfif")||h[i].contains(".png")||h[i].contains(".jpg")
  ||h[i].contains(".GIF")||h[i].contains(".PNG")||h[i].contains(".JPG")|| h[i].contains(".tif")||
   h[i].contains(".TIF")|| h[i].contains(".gif")|| h[i].contains(".JPEG"))
{
    l[i]=h[i].replaceFirst("[.][^.]+$", "");
    b="C:\\OCR\\Amharic\\"+l[i]+"";
    System.out.println(l[i]);
    File file = new File("C:\\OCR\\Amharic");
    file = new File(file, l[i]);
    try {
        if (!file.exists()) {
            file.createNewFile();
        }
        else
        {
            System.out.println("file exist!");
        }}
        catch (IOException ex) {
            Logger.getLogger(OCR_Recognition1.class.getName()).log(Level.SEVERE, null, ex);
        }}
        else
        {
            System.out.println("unable to create!");
        }

        System.out.println("created");
```

FIGURE 6
SAMPLE CODE STORING AMHARIC FILES

Then the app will access the Amharic trained data from the tesseract installation path and convert the image to editable text format and save it as text file with the name the same as the image file name. Here to maintain the layout similar to the actual image we used notepad++ as a tool. The following sample code shows how this happens. For converting scanned pdf files the system will convert each and every pages of the pdf file to image file and store it to certain

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

folder. Then it will automatically access each and every image file from that location and perform the conversion process in similar logic with that of image file (Figure 7).

```
String tesseract_install_path="C:\\Program Files\\Technology and innovation institute\\ocr-Application\\Tesseract-ocr\\tesseract";

String[] command =

{

   "cmd",

};

Process amh;

try {

   amh = Runtime.getRuntime().exec(command);

   new Thread(new SyncPipe(amh.getErrorStream(), System.err)).start();

   new Thread(new SyncPipe(amh.getInputStream(), System.out)).start();

   try (PrintWriter stdinn = new PrintWriter(amh.getOutputStream())) {

      String d="\""+tesseract_install_path+"\" \""+arrSplit[i]+"\" \""+b+"\" -l amh";

      stdinn.println(d);
      System.out.println(d);
      stdinn.close();

   }
```

FIGURE 7
SAMPLE CODE CONVERTING IMAGE TOE EDITABLE TEXT

Then it will store the result as separate text file as well as merged file format. The following shows the sample code (Figure 8).

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

```
String tesseract_install_path="C:\\Program Files\\Technology and innovation institute\\ocr-Application\\Tesseract-ocr\\tesseract";

String[] command =

{

   "cmd",

};

Process amh;

try {

   amh = Runtime.getRuntime().exec(command);

   new Thread(new SyncPipe(amh.getErrorStream(), System.err)).start();

   new Thread(new SyncPipe(amh.getInputStream(), System.out)).start();

   try (PrintWriter stdinn = new PrintWriter(amh.getOutputStream())) {

      String d="\""+tesseract_install_path+"\" \""+arrSplit[i]+"\" \""+b+"\" -l amh";

      stdinn.println(d);
      System.out.println(d);
      stdinn.close();

   }
```

FIGURE 8
STORING RESULT

CONCLUSION

The objective of the research was to develop OCR application that translates both Amharic and English based images and scanned pdf files into editable content. Accordingly attempt is made to investigate and develop OCR system that can recognize texts images in both Amharic and English language with better accuracy.

Generally, OCR application is one of the most important tools for easy extraction of text from an image without having the burden to type the text on a pc. The technology is new and current in which most worldwide languages are being incorporated. Since the technology must be adopted for many languages found in our country we recommend developers and researchers to work on various languages in addition to Amharic as well as to investigate more on how to enhance degree of accuracy of OCR applications for our languages.

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.

REFERENCES

Al-A'Ali, M., & Ahmad, J. (2007). Optical Character Recognition System for Arabic Text Using Cursive Multi-Directional Approach. *Journal of Computer Science, 3*, 549-555.

Chaudhuri, A., Gerlich, H.A., Jayaram, J., Ghadge, A., Shack, J., Brix, B.H., Hoffbeck, L.H., & Ulriksen, N. (2020). Selecting spare parts suitable for additive manufacturing: a design science approach. *Production Planning & Control, 32*, 670 - 687.

Das, S., & Banerjee, S. (2012). Pattern recognition approaches to japanese character recognition. In *Advances in Computer Science, Engineering & Applications* (pp. 83-92). Springer, Berlin, Heidelberg.

Das, S., & Banerjee, S. (2014). Survey of Pattern Recognition Approaches in Japanese Character Recognition. *International Journal of Computer Science and Information Technology, 5*(1), 93-99.

Dereje, T. (1999). *Optical character recognition of typewritten amharic text*. M.Sc. Thesis. Addis Ababa University. Faculty of informatics.

Meshesha, M., & Jawahar, C.V. (2007). Optical character recognition of Amharic documents. *African Journal of Information & Communication Technology, 3*(2).

Offermann, P., Levina, O., Schönherr, M., & Bub, U. (2009). Outline of a design science research process. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology* (pp. 1-11).

Sai Abhishek, B.V., Yamuna, K., & Anjali, T. (2021). Multilingual Translational Optical Character Recognition System for Printed Telugu Text. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1-5.

Wu, H., & Wang, K. (2020). Design of the Introduction Part of Information Processing and Machine Translation Course for Students Majoring in Computer Science and Technology. *2020 International Conference on Information Science and Education (ICISE-IE)*, 543-546.

Zayene, O., Touj, S.M., Hennebert, J., Ingold, R., & Amara, N.E. (2018). Multi-dimensional long short-term memory networks for artificial Arabic text recognition in news video. *IET Computer Vision, 12*, 710-719.

Received: 05-May-2022, Manuscript No. JMIDS-22-11926; **Editor assigned:** 09-May-2022, PreQC No. JMIDS-22-11926(PQ); **Reviewed:** 23-May-2022, QC No. JMIDS-22-11926; **Revised:** 27-May-2022, Manuscript No. JMIDS-22-11926(R)**; Published:** 31-May-2022

Citation Information: Beshah, T., & Asfawosen, A. (2022). Design and development of an ocr system that can convert both amharic and english based images and scanned pdf files into editable content. *Journal of Management Information and Decision Sciences, 25*(5), 1-11.