**IMAGE TO TEXT CONVERTER USING OPTICAL CHARACTER RECOGNITION TECHNOLOGY**

Joseph Odipo, Brandon Odhiambo, Silas Moracha, Teddysydney Odhiambo, Maxwell Mokua, Stephen Mbura, Dennis Gachemi and Anthony Oboch

School Of Computing And Informatics

Department Of Information Technology, Maseno University

CIT 208: Group Project

Dr. Roxanne Hawi

 Literature Review

**Image to Text Converter using optical character recognition system**

**1.0 Overview**

Today in the field of modern computing and artificial intelligence the need for an image to text converter using optical character recognition is a widely researched topic due to its numerous applications such as document scanning, licence plate identification, Business automation and handwriting recognition, Optical Character Recognition systems allow users to upload an image or scan a document and then by using its character recognition algorithms it will be able to read the text on the uploaded file and rewrite the words on the image in a text format that can stored on any text editor although there are some systems that can convert image to texts in real time through the use of  computer visions and Machine learning, despite developments and research in the field of deep learning algorithms and optical character recognition to help improve Optical Character Recognition systems, there are still numerous challenges in the journey to be able to produce an almost perfect image to text converter software. Some of the deep learning algorithms that are being developed to help better Optical Character Recognition software include: Convolutional-Recurrent Neural Network(CRNN), this approach identifies words using three steps; A standard convolutional neural network(CNN) where the first layer breaks the image into features and is divided into “Feature columns”, secondly these columns are fed into a deep-bidirectional long short term memory(LSTM)cell, which provides a sequence, identifying the relationship between the characters, finally the output of the LSTM cell is fed into a transcription layer which takes the character sequence, including redundant characters, and uses a probabilistic approach to clean the output, another deep learning algorithm is the Recurrent Attention Model(RAM) which is based on the idea that when the human eye is presented with a new scene, certain parts of the image catch its attention. The eye focuses on those glimpses of information first and obtains information from them. In the model, the image is cropped to different sizes around a common centre, and glimpse vectors are created with prominent features from each cropped version. These glimpse vectors are flattened and passed through a “glimpse network” based on visual attention. Glimpse vectors are then passed to a location network, which uses an RNN to predict the next part of the image to pay attention to. This location is the next input for the glimpse network. Gradually, the model explores additional parts of the image, each time performing backpropagation to see if the information from the previous glimpses is good enough to achieve a high level of accuracy and finally the Attention-Optical Character Recognition deep learning algorithm is whereby firstly, the model uses convolutional network layers to extract image features. It encodes these features to strings and passes them to an RNN, followed by an attention mechanism, borrowed from the Seq2Seq machine translation model. The attention-based decoder is used to predict the text in the input image. Optical character recognition technology has made advances in handwriting recognition, allowing it to recognize and convert handwritten text into machine-readable format. This can be useful in digitising handwritten notes, historical documents, and other handwritten materials, the most commonly used handwriting recognition system is the PTICAL character recognition, there are also Optical Character Recognition systems that have become more versatile, with the ability to recognize text in multiple languages. This allows users to convert text from a wide range of languages and scripts, making Optical Character Recognition technology more accessible and useful to a wider range of users. In this review we will examine some of the existing Optical Character Recognition systems, their shortcomings and some of the methods that we have employed to make our system better and much more accurate.

**1.1 Related work**

Photo scan is a mobile system that was developed by google to allow users to digitise their old photos using its Optical Character Recognition capabilities, it was designed to be simple and user friendly hence makes it easier to digitise their old photos for future use. The key advantage of photo scan is that it is able to recognise text in old photos and images that have been taken under less-than-ideal conditions for example faded images, torn or damaged images. Photo scan uses an advanced algorithm to digitise text in photos providing users with a reliable and accurate result. Another key advantage of photo scan is that it is able to integrate with google services such as google photos and google drive, making it easier for users to store their digitised photos in cloud hence making it possible to access them anywhere and share them with others. One of the main disadvantages of photo scan is that it tends to struggle when reading text from blurry images or if the text is written in unusual font, hence its accuracy might decrease, additionally photo scan capabilities may struggle with recognizing characters from other languages as it was designed to primarily identify texts in English. Photo scan also requires a reliable internet connection to function, as its Optical Character Recognition capabilities are a cloud-based service, this might be a problem for users who are located in areas with unstable internet connectivity or there is just no internet connectivity.

Adobe Acrobat Dc also has Optical Character Recognition capabilities, as it can convert scanned documents and images into searchable and editable PDF files. Accuracy is one of the key benefits of utilising Adobe Acrobat DC for Optical Character Recognition, the software can recognize very tiny and complicated fonts, producing extremely accurate Optical Character Recognition results, a built-in spell checker in Adobe Acrobat DC further ensures the accuracy of the typed text, also Adobe Acrobat DC provides a variety of Optical Character Recognition settings that lets its users to tailor their Optical Character Recognition procedure to their unique requirements. Users can specify the document's language, choose between editable and searchable PDFs, and change the picture resolution for the best Optical Character Recognition results, additionally Adobe Acrobat DC is a well-liked option for document management because it includes a lot of additional functions in addition to its Optical Character Resolution capabilities. For instance, it makes it simple to collaborate and share papers with others by enabling users to add comments, annotations, and digital signatures to PDF files, it also has a number of security features, including password protection and redaction tools. However, the cost of Adobe Acrobat DC for its Optical Character Recognition is one of its main disadvantages. It can be expensive for individuals and small organisations to use the program because it is a paid subscription-based service. Additionally, the Optical Character Recognition procedure can be tiresome and slow, particularly for larger texts and photos. Another further potential disadvantage of utilising Adobe Acrobat DC for Optical Character Recognition is that it could occasionally generate inaccurate results with documents that contain handwritten text or have poor image quality.

Google Drive is also another Optical Character Recognition software that is developed by google. Google Drive is a cloud-based file storage service that allows users to search for a text within images and scanned documents, this cloud-based service feature makes it possible to quickly find and access specific data contained within images and scanned documents, when an image or scanned document is uploaded to google drive, google drive automatically recognises the text and makes it editable and searchable. Google drive Optical Character Recognition capabilities are highly scalable making it possible for users to manage large collections of files as it can scan data up to 100 MB and images this is a huge advantage that google drive has over other optical Character Recognition systems as they tend to struggle when faced with the challenge of decoding texts for image in a large collection of data although google drive also struggles when dealing with large information but it is able to process the information faster when compared to other Optical Character Recognition systems that are available on the market. Language identification is another one of Google Drive's Optical Character Recognition’s most notable functions. With the aid of this capability, the program can automatically determine the language of a scanned image or document and use the proper Optical Character Recognition algorithm to increase its accuracy. This is particularly helpful for documents containing many different languages, which can be difficult for other Optical Character Recognition tools to process. However, one of google drive’s main disadvantage is that it might not be as accurate as some Optical Character Recognition systems which are only dedicated to Image text conversion because google drive’s main function is storage and retrieval of information rather than analysis of stored information, also like other Optical Character Recognition systems Google Drive’s Optical Character Recognition Capabilities are cloud based hence this might be a problem for user’s that are located in areas where there is no stable internet connectivity or just no internet connectivity.

Microsoft OneNote is also another example of a system that has Optical Character Recognition capabilities as it allows users to recognise text in images and scanned documents. OneNote's OCR features make it possible to digitise documents and make the data they contain easily accessible and searchable, which is one of their main advantages. In order to make it simple to locate specific information later on, OneNote, for instance, can detect text in receipts, business cards, and notes made on paper and make them searchable. This can be especially helpful for people and groups that need to keep track of a large amount of data. Optical Character Recognition capabilities in OneNote are also quite accurate, allowing users to trust the data in their notes with ease. Advanced algorithms are used by the application to identify text, even in scanned documents and photographs that may have been taken in less-than-ideal lighting circumstances. Users may rest assured that the data they enter into OneNote will be accurate and trustworthy as a result.  However, OneNote's Optical Character Recognition capabilities are limited, as one is required to pay so as to be able to access full Optical Character Recognition capabilities, it also has an accuracy of 70% depending on the quality of the document or image that has been uploaded which is quite low when compared to other Optical Character Recognition systems.

SimpleOCR is one of the first Optical Character Recognition Systems to enter the market, at the time of its launch simpleOCR was considered a pretty fast and accurate system but increase in demand of capabilities of Optical Character Recognition Systems such as support for multiple languages, led to numerous updates being made on the system this started making the system much more slower and clunky hence simpleOCR started to lose favour in the eyes of its users. SimpleOCR handwriting capabilities is also a paid subscription feature that its users can only get to use for a free fourteen day trial before they are required to pay so as to continue this feature.

Even though all these software’s have Optical Character Recognition capabilities each software has its own weakness and strengths, Google Drive is a cloud based software that scans uploaded documents on Google Drive, Photo scan is a mobile application that is used to digitise old photos and documents, Microsoft OneNote is a digital note taking application that can generate texts from uploaded images and documents, SimpleOCR which allowed users to also extract text from scanned images PDFs and photos, and Adobe Acrobat DC which  converts scanned documents and images into editable and searchable PDF files. Based on these comparisons Adobe Acrobat Dc is clearly the most advanced and accurate Optical Character Recognition Tool since it can recognize text in many different languages and maintain its layout and format, it is also very precise and fast, this is what sets Adobe Acrobat Dc apart from other competing Optical Character Recognition programs. With the help of sophisticated algorithms, Adobe's Optical Character Resolution technology effectively recognizes text and images. It can quickly and accurately convert scanned photos, PDF files, and other document types. Additionally, Adobe Acrobat DC has the ability to automatically identify and fix mistakes like twisted or skewed text, leading to more accurate conversions. Strong security features in Adobe Acrobat DC ensure that documents are shielded from unauthorised access. It provides several security features, such as password security, encryption, and digital signatures. To prevent mistakenly sharing confidential information, users can also redact sensitive information from papers. Users of Adobe Acrobat DC can modify the Optical Character Recognition settings to meet their own requirements. Depending on the type of document, the language, and the standard of the source material, users can choose from a variety of Optical Character Recognition settings.

**1.2 Literature Gap**

Indeed, literature demonstrates that these five Optical Character Recognition softwares have superior features such as extracting plain text from clear images , detecting the language of text  and be able to switch from one language to another due to multiple language they support, converting the texts to speech using well defined algorithms and many more appealing features. In spite of this , there is still a noticeable dearth in research pertaining to the five systems above -Photo scan Optical Character Recognition software, Google Drive Optical Character Recognition software, Adobe Acrobat DC Optical Character Recognition software, SimpleOCR and Microsoft OneNote Optical Character Recognition software – which is , they are cloud based hence they all need internet connection in order access and use their powerful features. This limitation poses a challenge because it limits a lot of people from using these softwares, because they are not aware of the existence of these softwares and the capabilities they have , these are the same people who might be interested in contributing to improving their performance .Thus , our proposed project , by using an Object Oriented Language in Java,  aims to close this literature gap by making it fully offline in order to be accessed in areas that have no internet access or connection, in addition to that , our proposed system focuses on enhancing the output of the Optical Character Recognition in a number of different image processing techniques ,our system will be able to give word suggestions where it detects violation of grammatical rules , by using Gaussian lowpass filtering ,It will be able to take blurry images, remove low frequency noise to make the image more soft , increase their clarity and light, refine them, preprocess them and threshold them to produce clear images with less noise on their backgrounds for easy text extraction. This way it is going to minimise most errors that the common Optical Character Recognition softwares in the market make. Last but not least ,  As computers can not be able to separate and name different objects in an image, our project implements Otsu’s optimum global thresholding to enable our optical character recognition system to be able to separate objects in an image thereby improving the overall output from the system. Also the image needs to go through morphological operations so as to remove gaps or outlying objects that can disrupt the image. Finally , this proposed project will be able to allow users to choose font and colour of the extracted text.By performing all these operations our Optical Character Recognition system output is greatly enhanced enabling us to produce quality text that is now cleaned using a spell checking mechanism as our systems output.

**1.3 Summary**

As described in the previous chapter, there exists a slightly high significant  literature  gap within the field of Photo scan Optical Character Recognition, the most common one being; many users being  unable to access Photo scan OCRs offline and use some of the features available in them since they are web based, which means they need constant internet access. Also many of the  Optical Character Recognitions have no ability to fully refine the image quality in order for the text to be extracted which leads to replacement or substitution of words or letters with other characters close to or not close to the same letters or words. Although these  highly sophisticated cloud based softwares are more advanced in extracting the text from the image,converting the extracted text to audio and have the ability to convert the text to other multiple languages they still lack some features. Therefore in order to bridge this literature gap to the contribution of this dialogue, our project proposes the integration of a software that is both accessible online and offline, to make suggestion for the unrecognised words and letter-that means having a tool like spell checker that is able to give you suggestion of what the unrecognised words might be. The other implementation that we desire to add is; the software being able to refine unclear images by applying different techniques discussed above, in order for the text to be extracted with high precision and minimum or no error,also to allow the users to change the font and colours from the extracted texts in real time rather than transferring the extracted text and editing it in other text based softwares such as microsoft word. The working of these tools and techniques have been discussed in detail in the subsequent chapter  and the underlying process they need to go through in order to achieve our listed desired goals.

**1.4 References**

Challenges facing current Optical Character Recognition software retrieved from

<https://theecmconsultant.com/disadvantages-of-optical-character-recognition/>

Deep Learning Algorithms retrieved from

<https://heartbeat.comet.ml/optical-character-recognition-using-deep-learning-techniques-1376605b022a>

Google Drive review retrieved from

<https://www.enolsoft.com/blog/google-drive-ocr.html>

Microsoft OneNote review retrieved from

<https://www.accuratereviews.com/ocr-software-reviews-list/microsoft-onenote/#:~:text=Microsoft%20OneNote%20Review&text=The%20OneNote%20OCR%20feature%20will,better%20with%20higher%20resolution%20images.>

Hand Writing recognition retrieved from

[https://www.researchgate.net/publication/343273822\_Handwritten\_Optical\_Character\_Recognition\_OPTICAL CHARACTER RECOGNITION\_A\_Comprehensive\_Systematic\_Literature\_Review\_SLR](https://www.researchgate.net/publication/343273822_Handwritten_Optical_Character_Recognition_OCR_A_Comprehensive_Systematic_Literature_Review_SLR)

Existing Optical Character Recognition software retrieved from

<https://www.makeuseof.com/tag/how-to-extract-the-text-from-images/>

 Photo Scan review retrieved from

<http://allaboutwindowsphone.com/flow/item/21624_Mini-review_Photo_Scan.php>

Similar projects being developed on Optical Character Recognition information retrieved from

<https://www.ijser.org/researchpaper/Literature-Survey-on-Recognition-and-Evaluation-of-Optical-Character-Recognition-OCR.pdf>

Additional information on similar projects being developed retrieved from

<https://www.irjet.net/archives/V6/i6/IRJET-V6I6736.pdf>

Adobe Acrobat Dc review retrieved from

<https://www.accuratereviews.com/ocr-software-reviews-list/acrobat-reader-dc/>

SimpleOCR review retrieved from

<https://perfectgeeks.com/simpleocr-review/>

Gaps and Tools Used retrieved from

[https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.diva-portal.org/smash/get/diva2:1566673/FULLTEXT01.pdf&ved=2ahUKEwjQmrytiJr9AhUMTKQEHU1PBsgQFnoECBMQAQ&usg=AOvVaw0Y6ofiTi3VjcKJqzdS\_56z](https://www.diva-portal.org/smash/get/diva2:1566673/FULLTEXT01.pdf)