Optical Character Identification

R. Mynavathi [1], S. Arvindh Prasadh[2], T. M. Divakar[3], K. Kishore[4], M. Kishore Manikandan[5]

[1]Department of Information Technology, Vellalar College of Engineering and Technology, Erode, Tamil Nadu, India.
Email: rpgmyna@gmail.com[1]

[2,3,4,5]UG Student, Vellalar College of Engineering and Technology, Erode Tamil Nadu, India.
Email: arvindhsugu@gmail.com[2], divakarthangaraj@gmail.com[3], kishore.rathika5304@gmail.com[4], kishov1999@gmail.com[5]

Abstract- Optical Character Identification (OCI) and text recognition applications are used commonly in business as well as in research. The real value is the effort and time that can be reduced by utilizing this type of application. There are a lot of applications, coding libraries and commentarial software for OCI in international languages. The current capacity to translate paper documents quickly and accurately into machine readable form using optical character recognition technology augments the opportunities in document sharing and storing. By having a scanner app it offers the workforce tool to get more organized and it will eliminate the time consuming manual process. The idea is to introduce a scanner app which is useful in capturing information on just everything from a document to presentation, receipts, business cards and a lot more. The converted text files take less space than the original image file. This will eliminate the manual processes making the information accessible anytime and from anywhere. The limitations of mobile device processor hinder the possible execution of computationally intensive applications that need less time of process. A framework of Optical Character Identification (OCI) on mobile device using server-based processing is proposed. It is inferred that the server-based mobile OCI obtains higher character recognition and format recognition accuracy than the standalone. The framework tries to overcome the limitation of mobile device capability process, so the devices can do the computationally intensive application more quickly.

Keywords: Optical Character Identification, labour reducing technology, Machine Intelligence, artificial intelligence, Advanced Systems.

I. INTRODUCTION

Optical Character Identification (optical character identifier, OCI) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast). It is widely used as a form of data 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. OCI is a field of research in pattern recognition, artificial intelligence and computer vision. Early versions needed to be trained with images of each character, and worked on one font at a time. Advanced systems capable of producing a high degree of recognition accuracy for most fonts are now common, and with support for a variety of digital image file format inputs.

II. OBJECTIVE

The main objective is to convert scanned images of typed or printed text to a text file. OC technology is used when recreating a similar document from a paper document to electronic form. The project has simplified yet another function which happens to be immensely useful like scanning by a scanning app turned as OCI mobile application. Once the text of the image has been identified it can be saved in the file storage. It can also be edited and shared.

A. Pattern Analysis and Machine Intelligence

George Nagy (2016) [2] proposed on Pattern recognition is generally categorized according to the type of learning procedure used to generate the output value. Supervised learning assumes that a set of training data (the training set) has been provided, consisting of a set of instances that have been properly labelled by hand with the correct output. A learning procedure then generates a model that attempts to meet two sometimes conflicting objectives: Perform as well as possible on the training data, and generalize as well as possible to new data (usually, this means being as simple as possible, for some technical definition of "simple", in accordance with Occam's Razor, discussed below). Unsupervised

learning, on the other hand, assumes training data that has not been hand-labelled, and attempts to find inherent patterns in the data that can then be used to determine the correct output value for new data instances. A combination of the two that has recently been explored is semi-supervised learning, which uses a combination of labelled and unlabelled data (typically a small set of labelled data combined with a large amount of unlabelled data). Note that in cases of unsupervised learning, there may be no training data at all to speak of; in other words, and the data to be labelled is the training data.

B. Pattern Recognition

J Mant (2018) [4] proposed the increasing prevalence of automated image acquisition systems is enabling new types of microscopy experiments that generate large image datasets. However, there is a perceived lack of robust image analysis systems required to process these diverse datasets. Most automated image analysis systems are tailored for specific types of microscopy, contrast methods, probes, and even cell types. This imposes significant constraints on experimental design, limiting their application to the narrow set of imaging methods for which they were designed. One of the approaches to address these limitations is pattern recognition, which was originally developed for remote sensing, and is increasingly being applied to the biology domain. This approach relies on training a computer to recognize patterns in images rather than developing algorithms or tuning parameters for specific image processing tasks. The generality of this approach promises to enable data mining in extensive image repositories, and provide objective and quantitative imaging assays for routine use.

III. EXISTING SYSTEM MODEL

In the running world there is a growing demand for the users to convert the printed documents in to electronic documents for maintaining the security of data. Hence the basic OCR system was on hardware and software but not that much accuracy. It was invented to convert the data available on papers. As OCR was in the form of hardware so that it could not manage to handle large amount of paper.

IV. PROPOSED SYSTEM

The proposed system is OCI ON SMARTPHONES (ANDROID APPLICATION). It is a character identification system that supports recognition of alphabets, numbers, and special characters. It gives a live pop up of all the data of what we scan. A fast response in translating large collections of image based electronic documents into structured electronic documents. This allows the users to directly share the text which is scanned by scanner. Data separation from large documents can be done. Most of the character recognition systems will be recognized through the input image with computer software. There is a large amount space require for computer software and scanner. In order to overcome this problem of computer and scanner occupying a large space, optical character identification (OCI) system, based on android phone is proposed. It is also used to avoid the performance bottlenecks of using standalone OCI techniques. OCI technology allows the conversion of image which is a scanned from a printed document, into text or any other information that user want using android mobile.

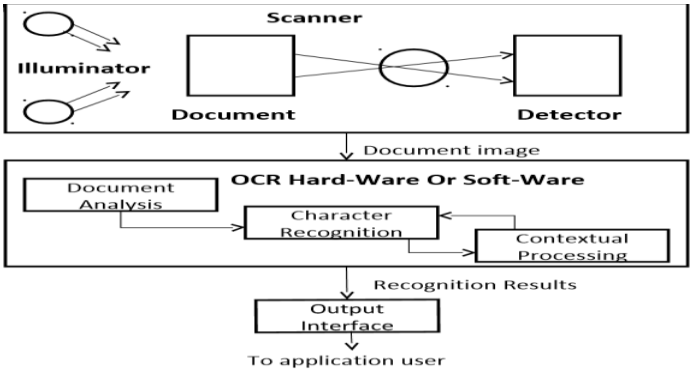

Fig 1 Architecture Diagram

V. MODULE DESCRIPTION

A. Scanning

A scanner is a device that captures images from photographic prints, posters, magazine pages, and similar sources for computer editing and display. Scanners come in and flatbed types and for scanning black-and-white only, or color. Very high-resolution scanners are used for scanning for high-resolution printing, but lower resolution scanners are adequate for capturing images for computer display. Scanners usually come with software, such as Adobe's Photoshop product, that lets you resize and otherwise modify a captured image. At first, the permission for camera access will be done. Auto focus function enables automatically. Flash light is an optional if needed which will be available in the check

box. The characters are scanned with the help of camera. The characters are scanned based on the dimensions of the data and the indentation.

B. Character Recognition

Text recognizer function is a predefined library file which is used for the character recognition. Based on the dimensions of the data each of the character is recognized. After the dimensions are recognized, each character of the data is compared with the text recognizer library file. After the comparison, the data is verified and popped up on the screen. Click the required data on the popped up data. When a character is identified, it is converted into an ASCII code that can be used by computer systems to handle further manipulations. Users should correct basic errors, proofread and make sure complex layouts were handled properly before saving the document for future use.

C. Storage and Sharing

The data will be fetched after the recognition process which will be displaying in the text view. Options for storing and sharing will be given in this application. Storage devices are one of the core components of any computing device. They store virtually all the data and applications on a computer, except hardware firmware. They are available in different form factors depending on the type of underlying device. For example, a standard computer has multiple storage devices including RAM, cache, and hard disk, as well as possibly having optical disk drives and externally connected USB drives. A storage device is any computing hardware that is used for storing, porting and extracting data files and objects. It can hold and store information both temporarily and permanently, and can be internal or external to a computer, server or any similar computing device. The fetched data can also be copied and pasted wherever required. The text can be directly shared to other android applications such as Gmail, What's app, Share it, etc., The fetched data can also be stored as a text document.

VI. CONCLUSION & FUTURE ENHANCEMENT

The automated entry of data by OCI is one of the most attractive; labour reducing technology. The recognition of characters by the system is very easy and quick. It is possible to edit the information of the documents more conveniently and can reuse the edited information as and when required. Training and recognition speeds can be increased greater and greater by making it more user-friendly.

In future, for the further enhancement of the work will be reading of handwritten which might be a difficult task. Reading of other languages can also be introduced.

REFERENCES

[1] Fischer S, Digital Image Processing: Skewing and Thresholding, Master of Science Thesis, University of New South Wales, Sydney, Australia, 2015.

[2] George Nagy (2016), "Twenty years of document image analysis is PAMI ". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 38-62, 2016.

[3] Giri K J , "Design and Implementation of a novel cognitive character recognition technique", International Conference on Signal Processing and Communication, pp. 225-229, 2015.

[4] Mant J (2018), "An overview of character recognition methodologies", Pattern Recognition vol. 19, pp. 425-430, 2018.

[5] Mori S, Suen C Y, Yamamoto K (2016), "Historical review of OCI research and development", Proc. IEEE vol. 80, pp. 1029-1058, 2016.