# PRINTED CHARACTERS TO DOCUMENT USING OCR – AN ANDROID APPLICATION

[1] Dr. R. Mynavathi, [2] Mr. P. Rajendran, [3] S. Arvindh Prasadh, [4] K. Bharath, [5] T. M. Divakar, [6] M. Kishore Manikandan

*[1 & 2]Associate Professor & [3, 4, 5 & 6]Final Year Students,*
*[1, 3, 4, 5&6]Department of Information Technology, [2]Department of Computer Applications,*
*Velalar College of Engineering and Technology, Thindal,*
*Erode, Tamilnadu, India, Pin Code: 638012.*
** Corresponding Author Email: arvindhsugu@gmail.com*

-------------------------------------------------------------------------------------------------------------

***Abstract:***
*Printed Paper documents today are in need of getting converted quickly and accurately into machine readable form using optical character recognition technology. This offers the opportunities in document sharing and storing. In the running world there is a huge demand for the users to convert the printed documents in to electronic documents which are in a high capacity for maintaining the security of data. By having this application it offers the workforce tool to get more organized. An application to capture data present in the mark sheet has been introduced. After the information is captured it will be stored in a spread sheet. It eliminates the time consuming manual process. Optical character recognition is an active research area that tries to boost and develop a system with the ability to extract and process text from images automatically.*

***Keywords**: Text Recognizer, Bitmap Conversion, Frame Separation, CSV - Comma Separated Values.*

## 1. INTRODUCTION

OCR is mostly used for data entry from which is helpful in printed paper data records, documents of passports, invoices, statements from banks, computerized receipts, credit or debit cards, e-mail, printouts of static- data, or any suitable documentation. It is a most common method of printed texts to digitized format so that they can be searched, edited, displayed on-line, stored more compactly and used in machine processes such as machine translation, (extracted) text-to-speech, cognitive computing, text mining and key data electronically. OCR is a research in the field of artificial intelligence, pattern recognition and Computer vision. Early versions of OCR needed to be trained with images of each and every character and needs to be worked on one font at a time. Advanced systems that are highly capable of producing a variety of digital image file format inputs with a support of high degree of recognition accuracy for most fonts. Optical character recognition is an active research area that helps a computer system to develop the ability of extracting the data and process the data from images automatically. These days there is a huge demand for storing information to a computer storage disk from the data available in printed or handwritten documents to later re-utilize this information by means of computers. A simple way to store information from these paper documents to a computer system is to scan the documents first and then store them as image files. Reutilizing of this information, from these image files is very difficult to read or query text. Therefore a technique to automatically retrieve and store information from image files in particular text is needed. Some of the major challenges needed to be pulled out and considered in order to achieve a successful automation. Some of the recent challenges are the font characteristics of the characters in quality of images and paper documents. Due to these challenges, computer system sometimes may not be recognized correctly. Thus there is a need of mechanisms of character recognition to perform Document Image Analysis (DIA) which will be immensely helpful to overcome these challenges and produces electronic format from the transformed documents in paper format.

## 2. OBJECTIVE

The main objective is to obtain electronic form of data that is scanned. OCR 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 application turned as OCR mobile application. An application to capture data present in the mark sheet and the recognized data will be stored in an excel sheet. Data Editing can be done.

## 3. RELATED WORKS

Various approaches for the detection, localization and classification of different characters in images have been proposed. The problem with the OCR system is about recognizing data from images of a mark sheet and storing it in the spread sheet. The solution requires comprehensive image processing that involves identification of numbers and pattern recognition. Some of the challenges in recognizing correct numbers are table variations in the mark sheets, marks location within the image, name and register number due to line spacing. Besides image variations, the amount of light and background can worsen the problem further. Big problem is less accuracy in fetching the data. Data Recognition takes time. As the data entry is done manually there might be a high chance of incorrect entry of data. Time consumes for manual entry.

Let us consider the various approaches used in the existing system.

### *3.1 Neural Networks*

This methodology implements the method of forward context feed for classification of characters. Backward-propagation algorithm is used to build the neural network. The preprocessing measures include standardization, scale and edge detection. The horizontal and vertical graph and component survey can tackle the character fragmentation problem.

### *3.2 Receipt Imaging*

Receipt imaging is broadly utilized as a part of n number of organizations applications to monitor financial and economical records and keep accounts of payments from heaping up. OCR simplifies information gathering in government offices and autonomous institutions, and analysis, among different procedures.

### *3.3 Tesseract Engine and Open CV*

This methodology uses adaptive threshold for highlighting the characters and remove the context. A component algorithm is first applied to the transformed binary image from the original image, in order to remove unnecessary image spaces. A special algorithm called Image Scissoring is used that makes use of Tesseract, an Optical Character Recognition Engine, which returns ASCII to the license number. The entire system was developed with the help of open CV.

### *3.4 MATLAB*

It provides an approach which is based on Sobel Edge's detection system and efficient morphological process. This approach is simplified by using the Surround Box method to separate both letters and numbers used in the number pad. The template matching methodology is used to recognize numbers and characters after the template is fragmented. The entire system was implemented with MATLAB.

### *3.5 Banking*

Another imperative use of OCR is in banking, without human intervention the cheques are processed. A cheque can be encoded with a machine where the framework eliminates the sum to be issued and the right measure of cash is exchanged. This innovation has been idealized for printed cheques, and is genuinely precise for handwritten checks diminishing the hold-up time in banks.

## 4. OCR PHASES

### *4.1 Pre – Processing Phase*

The aim of pre-processing is to eliminate noise in an image or undesired characteristics without missing any significant information. Pre-processing techniques demands on color, most likely grey-level or binary document images which contains text and graphics. Since color images are taken into account and it is computationally more expensive, grey images or binary images are utilized by most of the applications in character recognition systems. Pre-processing eliminates the inconsistent data and noise. It enhances the image and prepares the image for the next phases in OCR phases. The effectiveness and easiness of an image is enhanced by Pre-Processing phase which is the first phase that is to be processed in the next phases by converting the image to the suitable format. Therefore, the main issue in pre-processing phase decreasing the noise that causes the reduction in the character recognition rate. Thus, since pre-processing controls the successive phases for the suitability of the input, a primary stage prior to feature extraction phase is the pre-processing phase.

### *4.2 Segmentation Phase*

The major component and critical one of an Optical Character Recognition (OCR) system is the segmentation of text line from images. In general, Text segmentation from document image merges three segmentation and they are line segmentation, word segmentation and then character segmentation. Segmentation is the process of isolating or hiding text component from the image's background within an image. For reorganizing of the editable text lines from the recognized characters, firstly, the text lines are segmented, secondly, from the segmented line the words are segmented and then finally the characters are segmented. A major pre-processing phase in implementing an OCR system is document segmentation. It is the process of classifying a document images into different zones i.e., that each zone contains only one kind of information, such as a figure, a text, a halftone image or a table. In many cases, OCR highly depends on the accuracy of the page segmentation and the accuracy rate of systems that are related to the algorithm used.

### *4.3 Normalization Phase*

As a result of segmentation process the characters are segmented which are then moved through feature extraction phase, hence the segmented characters are minimized to a particular size depending on the algorithms which is used. The segmentation process converts the image in the form of m*n matrix for the normalization phase. These matrices are then commonly normalized by size minimization and eliminating the unnecessary or unwanted information from the image without missing any influential information.

### *4.4 Feature Extraction Phase*

Feature extraction is the process of extracting the pertinent features from alphabets or objects to build feature vectors. The feature vectors that are built is then utilized by classifiers to identify the input unit with objective output unit. To classify between dissimilar classes it becomes effortless for the classifier by having a glance at these features as it becomes fairly easy to determine. Several techniques for extracting features are proposed from the characters that are segmented. There is also a proposed one for directional chain code features and zoning and for handwritten numeral recognition considered to have a feature vector of length 100 and have achieved a high accuracy of recognition.

But, the feature extraction process takes consumption of time and complex. They have proposed and used horizontal/vertical strokes as the potential features for recognition and for handwritten numerals which has obtained a recognition accuracy of 90.50%. But, feature extraction method uses the process called thinning which leads to loss of some features. Statistical features are also taken into account as global features as the sub-images are usually extracted and averaged such as meshes. Initially, these statistical features are supplied for the recognition of machine printed characters. On the other hand, features of structural or topological are supplied to the character set that needs to be contemplated. Some of these features are concavities of number of holes in the characters and convexities in the number of end points etc.

### *4.5 Classification Phase*

OCR systems utilize the pattern recognition methodologies, in order to set each example to a predefined class. Classification is the procedure of distributing inputs with respect to detected information to their comparing class in order to create groups with homogeneous qualities, while segregating different inputs into different classes. Classification is conveyed to be on the premise of put away features in the feature space, for example, global features, structural features and so forth. It can be said that this classification takes several classes into account that isolates the feature space into the decision rule. Based on several agents the classifiers are choose, such as, n number of free parameters that are available for training set and so forth.

### *4.6 Post – Processing Phase*

It has been shown that people can read handwriting by context up to 60%. While pre-processing tries to eliminate the record in a by reducing the noise of the image, it might evacuate data which are critical, since the context data cannot be accessible at this stage. On the other chance might be accessible to a specific degree of semantic data, so that it would be able to contribute a solid measure to the precision of the OCR stages. On the other hand, the whole OCR issue is for deciding the context data of the image which is saved after the scanning process. In this way the incorporation of shape and context data in all the phases of OCR frameworks is vital for understandable upgrades in recognition rates. This is done in the Post processing stage with an input to the early phases of OCR. The least complex method for manipulating the context data for reducing the minor errors of the OCR frameworks is the usage of a dictionary. The fundamental thought is to consider the dictionary to spell check the OCR yield. This gives a few distinct options for the yields of the recognizer. The dictionary is nothing but the pre defined library files. The dictionary is also called as library files.

## 5. PROPOSED SYSTEM

The proposed system aims at detecting the marks of a mark sheet and recognizes its characters in an efficient and cost effective way.

The methodology followed in extracting the marks from the mark sheet using OCR is stated in the following steps.

a. Capture the image of the mark sheet.
b. Perform conversion of image to bitmap.
c. Recognition of characters from the bitmap.
d. Elimination of other characters other than numbers.

## ARCHITECTUTRE DESIGN

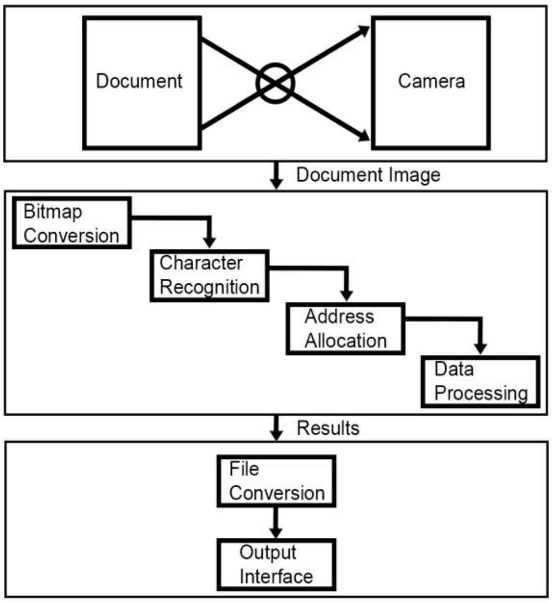

Fig.1.Block Diagram for implementation of the Proposed System

The proposed system comprises of three basic modules which are explained below.

### *5.1 Image Capturing / Accessing From Gallery*

This is the first and the most important stage of the system. This is the stage where the position of the marks is determined. The input at this stage is an image of the mark sheet and the output is the marks of the mark sheet. Here the image is converted into bitmap. At first, the permission for camera and gallery access will be done. Auto focus function gets enabled. Using camera the mark sheet's image is captured, the image is cropped over the area of marks obtained. In the other way, the images which we have captured already can be accessed from the gallery and it can be cropped.

### *5.2 Bitmap Conversion*

The image of the mark sheet captured is given as input to the Bitmap Conversion module. This is the stage, where the image of the mark sheet are mapped out and segmented into several frames. For segmenting the frames, the bitmap library file was used. The image is divided into several frames, each containing one isolated character. Each frame's address has been collected from the bitmap in order to be used in the recognition process. The values from the bitmap cannot be fetched directly so the address of the each frame is used for fetching the values.

### *5.3 Data Recognition*

For recognizing the characters, we have to use the text recognizer library file. This text recognizer is a pre defined library file. Each character recognized will be fetched and compared with the text recognizer library file and once they both match then the character will be stored. Most of the character recognition systems use this concept which is easier and an effective one. This recognition is done based on the input and this library file doesn't recognize multi languages. Offline android apps don't involve multi language recognition and it is not implemented. The characters from the bitmap are recognized and the address of each frame is stored in a sparse array using text block library file. The recognized data is fetched from the each frame's address and finally stored in a String Array. To store the next person's marks again the main activity is called and it goes to the main page. The process gets repeated till the user wants.

### *5.4 File Export*

The output is to display it in a Spread Sheet which is a file. The format used here is .CSV. A CSV is a comma separated values file, which is used to save data in a tabular format. CSVs look like a spread sheet with a .csv extension. Many spread sheet program can be used with .CSV file, such as Microsoft Excel or Google spread sheets. They help us to export a high volume of data to a more concentrated volume. They also serve two other primary functions, CSV files are plain text files, making them easier for the website developer to create. As they are plain text, they are easier to import into a spread sheet, regardless of the specific software you are using. The fetched data will be appended in a String array by using commas so that when the output is pushed as a .csv file they gets stored in the spread sheet. The file can be accessed from the internal storage and option for sharing is given. The file can be edited and stored again.

## 6.CONCLUSION

In this paper, an approach called printed characters to document using OCR through an android application has been presented to provide an efficient, cost effective system for detecting and recognizing the marks from the mark sheet. Although the OCR's text recognition is of high accuracy, many more practical applications of OCR can be done which will be useful in day today life. As a future work we are planning to use OCR for daily personal use. We are planning to make mobile devices incorporate with OCR in one OCR system. An automated book reader or receipt trackers are some of our future OCR based applications. The Image Capturing, Bitmap conversion, Data Recognition, Frame Separation and File Export has been done effectively using the Android Studio tool and its respective library files. This recognition has been trained using the android application with the help of the datasets collected and tested which acquires a better results and overall accuracy compared to the previous approaches.

### *REFERENCES*

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

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

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

[4] *Mandaviya K., Chaudhuri A., Badelia P., Ghosh S.K. (2018) Optical Character Recognition Systems. In: Optical Character Recognition Systems for Different Languages with Soft Computing. Studies in Fuzziness and Soft Computing, vol 352. Springer, Cham.*

[5] *Sabu A.M. and Das A.S. (2018), "A Survey on various Optical Character Recognition Techniques," 2018 Conference on Emerging Devices and Smart Systems (ICEDSS),pp. 152-155.*

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

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

[8] *D. Berchmans and S. S. Kumar (2014), "Optical character recognition: An overview and an insight," 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Kanyakumari, pp. 1361-1365.*

[9] *Patel C, Patel A, Patel D. Optical character recognition by open source OCR tool tesseract: A case study. International Journal of Computer Applications. 2012 Jan 1;55(10).*