OPTICAL CHARACTER RECOGNITION(OCR) TEXT DETECTION USING TESSERACT

Gaytri Sirskar[1], Mrunali Wande[2] , Bhagyashree Bobde[3] , Dhanraj Morkhade[4] , Nandan Pokale[5], Prof. S. R. Gudadhe[6]

[1,2,3,4,5] Students of Final Year, Sipna College of Engineering and Technology, Amravati, Maharashtra, India
[6]Assistant Professor, Department of Computer Science and Engineering, Sipna College of Engineering and Technology, Amravati, Maharashtra, India

---***---

Abstract – *Digitalized data is required in the modern day to enable speedier task completion and processing. Extracting the text from the images is the most effective technique to digitalize them. Many text identification task including image text recognition and optical character recognition can be used to process text. The technology of optical character recognition(OCR) was used to transform printed text into editable text. In a variety of applications, OCR is very helpful and popular approach. Text preparation and segmentation techniques can influence OCR accuracy. It is a technology that recognizes text within a digital image.*

These days, there is a huge need for information that may be found on paper, such books or newspapers. Information can currently be stored by scanning the desired text, but this method just saves the data as an image that cannot be processed further. For example, text recorded in scanned photos cannot be read line by line or word by word; we would have to completely redo the language included in these images before we could use them again. Text detection from papers when text is integrated in intricately colored document images is an extremely difficult problem. Many possible users would like to extract text from documents, archive documents, and other images. The user needs optical character recognition (OCR) because of this. Its goal is to identify textual areas in the document and distinguish them from the graphical section. obtaining data straight from application forms and greatly reducing time.

This study explains the basic ideas behind the OCR, including feature extraction strategies, picture preprocessing approaches, and recognition algorithms. It highlights significant turning points and innovations as it examines the development of OCR technology from early character recognition systems to contemporary deep learning-based techniques.

***Key Words*: Optical Character Recognition, Tesseract, Python, Django, Image Preprocessing, OpenCV**

1.INTRODUCTION

The need for effective and quick document processing and digitization has increased in an increasingly digital environment. Manual data entry and document handling techniques are time-consuming and error-prone, which reduces productivity and obstructs the smooth flow of information. This environment has completely changed with the introduction of optical character recognition(OCR) technology, Which makes it possible to automatically extract text from photographs and expedite the digitalization of documents. To fulfill the needs of contemporary information processing requirements, OCR systems speed and efficiency are still vital components.

OCR is commonly utilized in banking, where it can process demand drafts and checks without the need for human intervention. With the use of a smartphone camera, one can instantaneously scan the writing on a demand draft or check, transferring the exact amount of money. This method is fairly accurate for handwritten demand drafts or checks as well, however signature verification is necessary. It is almost perfected for printed demand drafts or checks. A notable trend toward digitizing paper documents has also emerged in the legal sector. Documents are being digitized to reduce storage requirements and do away with the need to go through bins of paper files. By enabling text searching for documents, OCR streamlines the process even further by making it simpler to find and manipulate them within the database. Legal practitioners can now quickly and easily search through a vast electronic document repository by only entering a few keywords. Numerous other industries, like as education, banking, and government organizations, heavily rely on OCR.Our technology is ready to enable not just faster transitions to paperless settings but also higher levels of data accessibility and accuracy, from the legal to the financial, healthcare and government sectors.

An eye can recognize, view and extract text from images, but person's brain must analyze any text that the eye detects or extracts. Naturally, OCR technology is still not as sophisticated as human talent. The quality of the input that the eye reads directly affects hoe well the brain functions when it comes to text recognition in humans. Numerous issues and difficulties may arise during the planning and execution of a computarized optical character recognition system. For instance some figures and letters differ just enough frm one another for computers to accurately identify them and separate them from one another. For instance, computers could find it difficult to distinguish between the numbers ”0” and ”o”, particularly when this characters are incorporated into a loud, extremely black background.

2. LITERATURE REVIEW

Proir OCR technology research has mostly concentrated on managing multilingual text, increasing accuracy, and connecting with different web application frameworks. Tesseract OCR has become a well-liked option because of its ongoing development and open-source nature. Research has demonstrated how well Tesseract OCR works to reliably extract text from images in a variety of languages.

- Although character recognition is not a brand-new issue, its origins can be found in earlier computer-related technologies. The first optical character recognition (OCR) systems were mechanical devices rather than computers, and they could recognize characters with relatively low precision and very slow speed. One of the earliest examples of modern OCR is the reading and robot GISMO, created in 1951 by M. Sheppard [1]. GISMO is able to read individual words on a printed page as well as musical notation. It can only identify 23 characters, though. A typewritten page could also be copied by the machine. In 1954, J. Rainbow invented a device that can read one capital typewritten English character each minute. The faults and poor recognition speed of the initial OCR systems drew criticism. As a result, during the 1960s and 1970s, little research was done on the subject. The only developments were made to huge enterprises and government organizations, such as banks, newspapers, airlines, and so on.

- It was decided that three OCR typefaces should be standardized in order to make the process of recognition for OCR easier due to the difficulties involved. As a result, in 1970, ANSI and EMCA established OCRA and OCRB, which offered recognition rates that were generally acceptable[2].

- A significant amount of study has been done on OCR over the last thirty years. As a result, multilingual, handwritten, omni-font, and document image analysis (DIA) have emerged [2]. Even after all of this research, the machine's reliability in reading text is still far behind that of a human. Therefore, current OCR research aims to increase OCR speed and accuracy for texts authored in a variety of styles and printed in unrestricted situations. There isn't any commercial or open source software available for difficult languages like Sindhi or Urdu, for example.

- “Extracting Text from Image Document and Displaying ITS Related Information”, K. N. Natei[3], journal of Engineering Research and Application: Image text is text that has been written or inserted into an image in a variety of formats. Captured photos, scanned papers, periodicals, newspapers, posters, and other materials can all contain visual text. The representation, description, and transmission of information that these image texts—which are widely available these days—help people with communication, problem-solving, availability, the creation of new job types, cost-effectiveness, productivity, globalization, and cultural gaps, among other things. If these picture documents' content were transformed to text, it would be more accessible and efficient. Text extraction is the method by which image text is transformed into plain text so that a computer can identify its ASCII characters.

- “Text Recognition using Image Processing”, International journal of Advanced Research in Computer Science: Text recognition by optical character recognition (OCR) uses a computer system intended to convert images of typewritten text into machine-editable text or photographs of characters into a common encoding scheme. OCR was first established as a subject of study in computer vision and artificial intelligence. Text recognition is employed in real-world applications where we wish to extract information from text written images, such as in post offices, banks, colleges, and other formal tasks where a lot of data must be typed.

3. ARCHITECTURE

Tesseract is an open-source optical character recognition (OCR) engine. This flowchart shows the text extraction procedure [1]. Tesseract is used to extract text from photos and translate it into machine-readable programming.

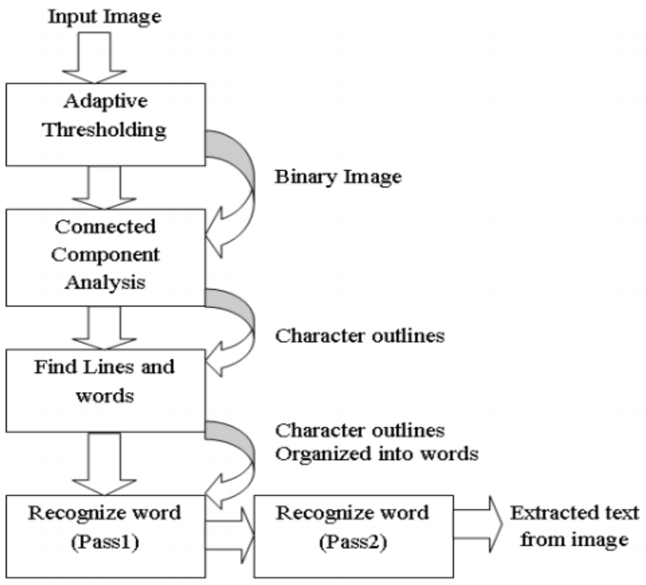


Fig-1 : Architecture of Tesseract OCR

The flowchart shows the several steps in this process:

The input image is the first picture from which you want to extract the text.Adaptive Thresholding: This phase creates a binary picture from the supplied grayscale image. It is simpler to differentiate text from background in a binary image since it simply has black and white pixels. Connected Component Analysis: In this case, the image is examined to determine and distinguish connected components, which are collections of black pixels that correspond to the text's letters . Character Outlines: Character shapes are refined by creating outline around the connecting components. Locate Lines and Words: The program locates and divides text lines as well as specific words inside those lines. Name the Word (Pass 1 & 2): To process the detected words, Tesseract uses a two-pass recognition algorithm. It makes an effort to identify each word separately on the first pass. High confidence word recognition is utilized to teach the OCR engine, increasing the accuracy of the second run. In the second pass, Tesseract applies the knowledge gained from the first run to improve its analysis of the remaining words. Extracted Text from Image: The output containing the identified text is provided at this point.

4. METHODOLOGY

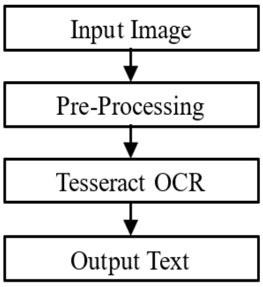


Fig-2 : Steps to generate output text from input text

In order to design,construct,an evaluate the suggested OCR system integrated inside the Django framework,along with the integration of a prediction data, a methodical set of procedures is used in this study.

4.1 System Design

The OCR systems architecture is painstakingly constructed within the Django framework during the first phase of system design.For text extraction, language selection, and accuracy analysis intregration, this means creating modular components that can be added to the system to guarantee smooth operation and communication.

4.2 Data Collection

After that a varied dataset is gathered for training and testing that includes pictures with text in Hindi, Marathi, English. The OCR system is trained on this dataset, which is also used to access the system's performance in variance language situations.

4.3 Preprocessing

The collected photos are preprocessed before being used in order to improve text visibility and maximize OCR accuracy.

To ensure the best conditions for text extraction, this preprocessing processes may include contrast modification, noise reduction, and image enhancement.

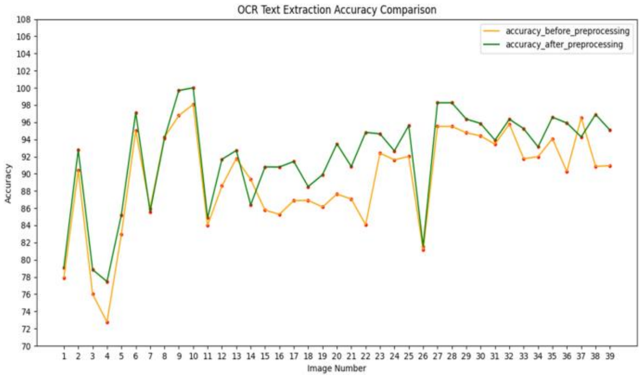


Chart-1 : Accuracy of Image Preprocessing

4.4 Implementation

Tesseract OCR integration with Django for web application development constitutes the fundamental implementation phase. This includes creating strong functions for language selection, text extraction, and chatbot integration while combining the best features of both systems. In addition, language selection capability is designed to improve user experience by enabling users to designate their preferred language for text extraction.

4.5 Testing and Evaluation

To determine its correctness and effectiveness in a range of linguistic contexts, the developed system is put through a thorough testing and assessment process. The OCR systems performance is tested using a variety of photos with text in various languages.

5. DESIGN

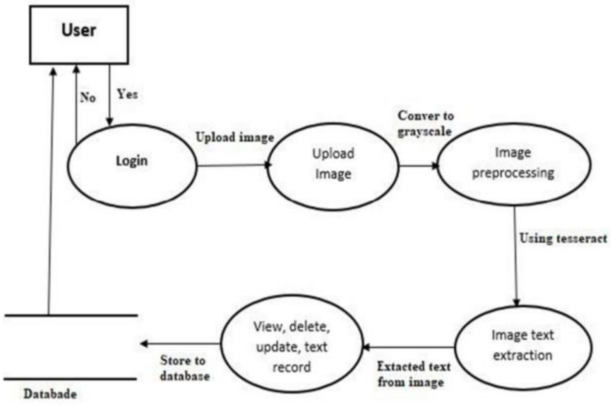


Fig-3: Block Diagram

The process's steps are as follows, as depicted in the flowchart:

- User: The user engages with the system to start the procedure. Logging into a website or application may be necessary for this.

- Upload Image: The user uploads an image file to the system by choosing it from their device.

- Yes/No Decision: The user may be asked by the system if they would like to preprocess the image. Resizing the image, changing its format, or adding filters are examples of preprocessing.

- Convert to Grayscale (Optional): The image is converted to grayscale if the user selects "yes" in the preceding stage. This can help the image's file size decrease and improve its database storage efficiency.

- Image Upload: The database receives the submitted image. The picture data is kept in the database, along with any other details about the image, like the filename, size, and upload date.

- Store to Database: The database contains a copy of the uploaded picture together with any further details about it.

- View, Delete, Update, Text Record: The user has the ability to view, delete, or update the database-stored details about the image. A caption, tags, or other descriptive text may be a part of this information.

6. PROPOSED SYSTEM

The objective behind the developed model as it is presented is to take in a lot of photographs of people's identity documents and categorize them into groups, like licenses and passports. The text extraction module is applied to the photos once they have been classed. From the categorized photos, the text data is taken out. Following the extraction of the credentials from the pictures, kept within the database. Extracting Texts The Tesseract OCR package is used to implement text extraction. It consists of the command line tool Tesseract and the optical character recognition (OCR) engine libtesseract. The Long Short-Term Memory (LSTM) based OCR engine in Tesseract is a novel neural net that specializes in line recognition while also recognizing character patterns. The recurrent neural network's building block is the LSTM network. Python-Tesseract is an OCR (optical character recognition) tool for text extraction.

7. TYPES OF OPTICAL CHARACTER RECOGNITION SYSTEMS

In previous years, research on OCR has been conducted in a wide range of directions. This section addresses the various kinds of OCR systems that developed as a consequence of these studies. These systems can be grouped according to factors like font limitations, character connection, and image acquisition mode. Fig. 4 classifies the system of character recognition.

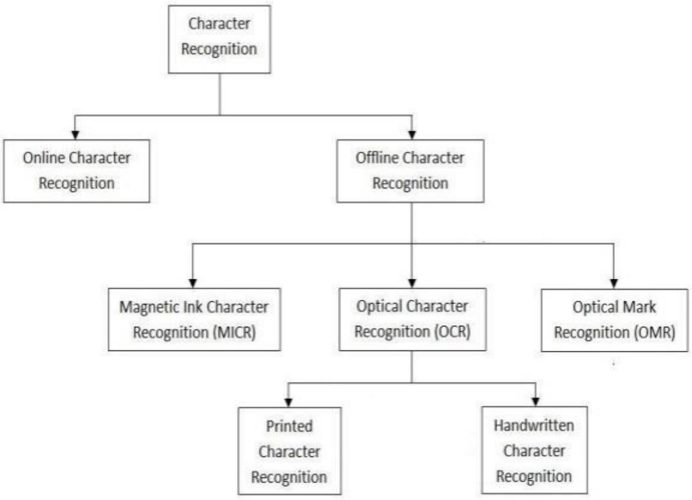

Character Recognition
Online Character Recognition
Offline Character Recognition
Magnetic Ink Character Recognition (MICR)
Optical Character Recognition (OCR)
Optical Mark Recognition (OMR)
Printed Character Recognition
Handwritten Character Recognition

Fig-4: Types of OCR System

OCR systems fall into two categories: machine printed character recognition and handwriting recognition, depending on the type of input. Because characters in the former case are typically of a uniform dimensions and character placements on the page are predictable [3].

Character recognition in handwriting is a particularly difficult task since various users have distinct writing styles and utilize diverse pen movements for the same character. These systems can be further subdivided into online and offline systems. While users are writing the character, the former is done in real time. Because they may record temporal or time-based data, such as speed, velocity, the quantity of strokes created, the direction in which the strokes are written, etc., they are less complicated. Additionally, since the pen's trace is only a few pixels broad, thinning procedures are not necessary. The bitmap input used by offline recognition systems is static data. As a result, performing recognition is quite challenging.

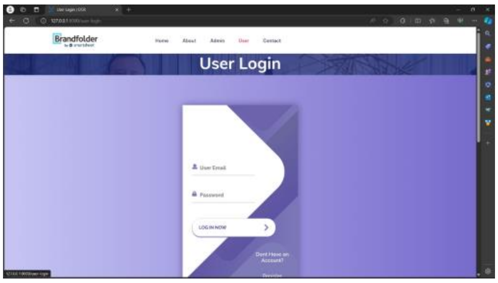
Due to their ease of development, high precision, and compatibility with tablets and PDAs, online systems have become widely available [4].

8. APPLICATIONS OF OCR

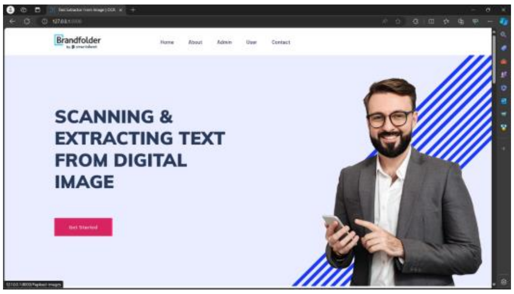
OCR makes a great deal of beneficial applications possible. OCR was first applied in mail sorting, bank check reading, and signature verification [5]. Organizations can also utilize OCR for automated form processing in locations where a large amount of data is available in printed form. OCR is also used for automatic number plate recognition, pen computing, passport validation`, utility bill processing, and other purposes [6]. OCR is also helpful for making text easier to read for blind and visually impaired individuals[7].

9. SNAPSHOTS

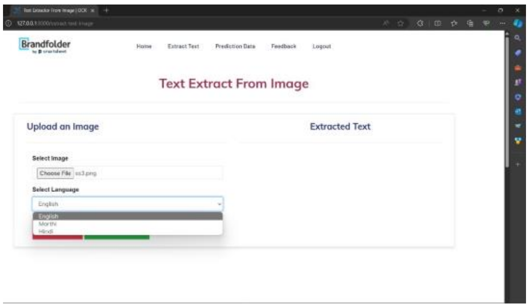
A. User Login Page



B. Home Page



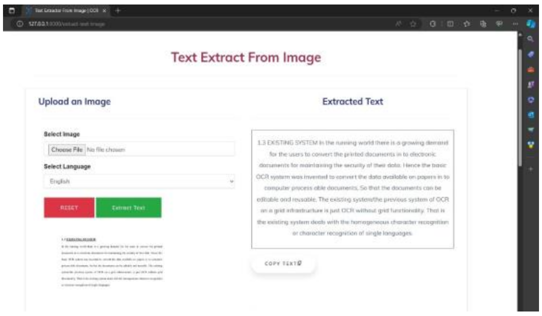
C. Upload Image



Select Image: ss3.png

Select Language: English

D. Extracted Text

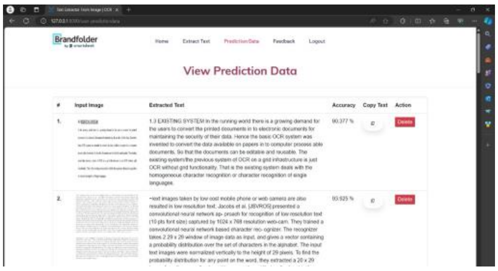


Select Image: No file chosen

Select Language: English

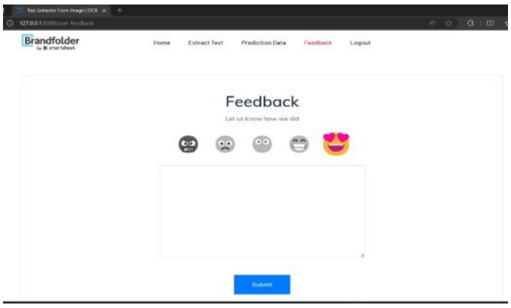
1.3 EXISTING SYSTEM In the running world there is a growing demand for the users to convert the printed documents in to electronic documents for maintaining the security of their data. Hence the basic OCR system was invented to convert the data available on papers in to computer process able documents. So that the documents can be editable and reusable. The existing system/the previous system of OCR on a grid infrastructure is just OCR without grid functionality. That is the existing system deals with the homogeneous character recognition or character recognition of single languages.

E. Accuracy Analysis

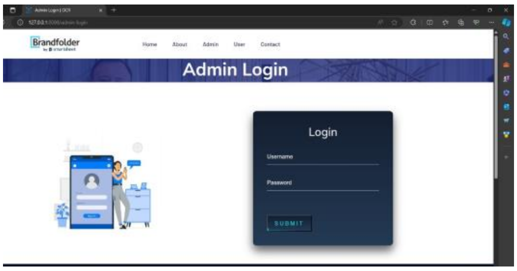


#	Input Image	Extracted Text	Accuracy	Copy Text	Action
1.		1.3 EXISTING SYSTEM In the running world there is a growing demand for the users to convert the printed documents in to electronic documents for maintaining the security of their data. Hence the basic OCR system was invented to convert the data available on papers in to computer process able documents. So that the documents can be editable and reusable. The existing system/the previous system of OCR on a grid infrastructure is just OCR without grid functionality. That is the existing system deals with the homogeneous character recognition or character recognition of single languages.	90.377 %		Delete
2.		~text images taken by low cost mobile phone or web camera are also resulted in low resolution text. Jacobs et al. [JSVR05] presented a convolutional neural network ap- proach for recognition of low resolution text (10 pts font size) captured by 1024 x 768 resolution web-cam. They trained a convolutional neural network based character rec- ognizer. The recognizer takes 2 29 x 29 window of image data as input, and gives a vector containing a probability distribution over the set of characters in the alphabet. The input text images were normalized vertically to the height of 29 pixels. To find the probability distribution for any point on the word, they extracted a 20 x 29	93.925 %		Delete

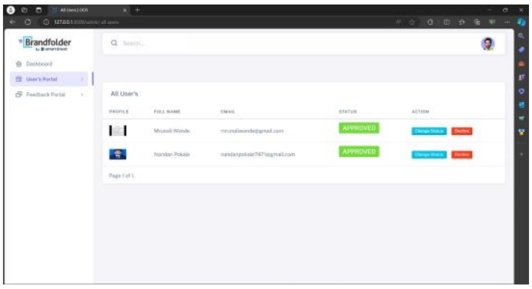
F. Feedback page



G. Admin Login Page



H. User Details



All User's

PROFILE	FULL NAME	EMAIL	STATUS	ACTION
	Mrunali Wande	mrunaliwande@gmail.com	APPROVED	Change Status Decline
	Nandan Pokale	nandanpokale7471@gmail.com	APPROVED	Change Status Decline

Page 1 of 1

10. FUTURE WORK

An OCR systems future score is predicted by taking into account a number of variables, such as technological developments, continuous R&D initiatives and prospective shifts in user needs and preferences. Even though it's difficult to assign a precise numerical score for an OCR systems future performance, we may make educated guesses about possible break through and areas for work that might boost the systems overall effectiveness . Taking into account the following factors.

The OCR technology has advanced. Future versions of OCR systems are probability going to gain for better algorithms, more advanced image processing methods, and more language support as OCR technology gets better. This may result in improved handling of various text layouts and fonts, increased accuracy rates and quicker processing times.

Using machine learning(ML) and artificial intelligence(AI) Methods in OCR systems has the protencial to increase performance and accuracy. Future optical character recognition(OCR) systems could becomes more accurate and efficient, especially in noisy or complicated contacts, by utilizing AI models for text interpretation and character recognition.

Futures OCR systems are anticipated to provide increased language support, incompassing a wider range of languages and dialects, the demand for multilingual OCR solutions is driven by globalization. As a result user would be able to realiably and accurately extract text from images in their own languages.

User experienced we probably will be given top priority in future OCR systems, which will likely include more user-friendly interfaces, smooth work flow integration and improved accessibility features to make OCR technology more accessible to people with disability might involve support for assistive technologies like screen reader and voice commands.

11. CONCLUSION

The application which is proposed in this report can be effective which will save a lot of time and money of common people. The primary goal in creating this program was to automate the process of creating a system that can reliably and effectively recognize text from pictures, text images, and scanned documents so that it can be used again at a later time. This article aims to provide an overview of What is the purpose of the project, what technologies are being utilized, what database is being used, why the application needs to be developed, and what methods are needed to make the project work.

REFERENCES

[1] Satti, D.A., 2013, Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach. MS thesis report Quaid-i-Azam University: Islamabad, Pakistan. p. 141.

[2] Mahmoud, S.A., & Al-Badr, B., 1995, Survey and bibliography of Arabic optical text recognition. Signal processing, 41(1), 49-77.

[3] Bhansali, M., & Kumar, P, 2013, An Alternative Method for Facilitating Cheque Clearance Using Smart Phones Application. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 2(1), 211-217.

[4] Qadri, M.T., & Asif, M, 2009, Automatic Number Plate Recognition System for Vehicle Identification Using Optical Character Recognition presented at International Conference on Education Technology and Computer, Singapore, 2009. Singapore: IEEE.

[5] "Extracting text from image document and displaying its related information", K.N. Natei journal of Engineering Research and Application (ISSN : 2248-9622, Vol. 8, Issue5 (Part -V) May2018

[6] K. Gaurav and Bhatia P. K., "Analytical Review of Preprocessing Techniques for Offline Handwritten Character Recognition", 2nd International Conference on Emerging Trends in Engineering & Management, ICETEM, 2013.

[7] Shen, H., & Coughlan, J.M, 2012, Towards A Real Time System for Finding and Reading Signs for Visually

Impaired Users. Computers Helping People with Special Needs. Linz, Austria: Springer International Publishing.

[8] Bhavani, S., & Thanushkodi, K, 2010, A Survey On Coding Algorithms In Medical Image Compression. International Journal on Computer Science and Engineering, 2(5), 1429-1434.

[9] Bhammar, M.B., & Mehta, K.A, 2012, Survey of various image compression techniques. International Journal on Darshan Institute of Engineering Research & Emerging Technologies, 1(1), 85-90.