ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Approach for Optical Character Recognition on Grid Infrastructure Using Kohonen Neural Network

Shantanu Chaudhary*, Swarnima Garg, and AbhinavBehera
Student, School of Computer Science and Engineering
VIT University. Vellore, India

Sathyaraj R
Assistant Professor (Senior),
Dept. of Software Systems
(School of Computer Science and Engineering),
VIT University, Vellore, India

Abstract: There is developing interest for the product frameworks to perceive characters in PC framework when data is looked over paper archives as we realize that we have number of daily papers and books which are in printed arrange identified with various subjects. This procedure is likewise called document image analysis. To viably utilize Optical Character Recognition for character acknowledgment so as to perform Document Image Analysis, we are utilizing the data in Grid arrange. For document processing archive preparing we require a product framework called character recognition system. Along these lines the need is to create character acknowledgment programming framework to perform Document Image Analysis which changes records in paper organization to electronic arrangement. Now and then in this record handling we have to prepare the data that is identified with dialects other than the English. With the help of kohonen neural network training handwritten character recognition is also done.

***Keywords*:** perceive character, document image analysis, Grid arrange, electronic arrangement, neural network, handwritten character recognition.

I. INTRODUCTION

In today's world, there is a developing interest for the clients to change over the printed records into electronic archives for keeping up the security of their information. Along these lines the need is to create character acknowledgment programming framework to perform Document Image Analysis which changes records in paper organization to electronic arrangement. For this procedure, there are different systems available. Among each one of those methods we have picked Optical Character Recognition as primary crucial procedure to perceive characters[1].The essential goal is to accelerate the procedure of character acknowledgment in record preparing. Therefore the framework can prepare tremendous number of archives with-in less time and thus spares the time[2]. From now on the major OCR structure was created to change over the data available on papers into PC get ready fit records, so that the chronicles can be editable and reusable. The current framework/the past arrangement of OCR on a matrix foundation is only OCR without lattice usefulness. That is the current framework manages the homogeneous character acknowledgment or character acknowledgment of single dialects. The extent of Optical Character Recognition on a lattice foundation is to give a proficient and improved programming device for the clients to perform Document Image Analysis, report preparing by perusing and perceiving the characters in research, scholastic, administrative and business associations that are having expansive pool of recorded, filtered pictures[4]. Our proposed framework is OCR on a matrix foundation which is a character acknowledgment framework that backings acknowledgment of the characters of numerous dialects[5]. This component is the thing that we get organize establishment which takes out the issue of heterogeneous character affirmation and support distinctive functionalities to be performed on the record.. In this unique situation, Grid foundation implies the framework that backings gathering of particular arrangement of dialects. Subsequently OCR on a matrix foundation is multi-lingual. The advantage of proposed framework that beats the disadvantage of the current framework is that it underpins different functionalities, for example, altering and looking. It likewise includes advantage by giving heterogeneous characters' acknowledgment[6]. The numerous functionalities incorporate altering and seeking as well whereas the current framework underpins just altering of the archive[7].The change of paper reports into electronic arrangement is an on-going assignment in huge numbers of the associations especially in Research and Development (R&D) region, in vast business endeavors, in government organizations, so on [11]. From our issue proclamation, we can present the need of Optical Character Recognition in portable electronic gadgets, PDAs, advanced cameras to gain pictures and remember them as a piece of face acknowledgment and approval. Some the areas where our venture can be applied are autonomous number plate recognition, invoices, screenshots, ID card, driver license[13].

Since our character acknowledgment depends on a Kohonen Neural Network(KoNN) foundation, it expects to perceive numerous heterogeneous characters that have a place with various all inclusive dialects with various textual style properties and arrangements[15].Regardless of the measure of archives and the sort of characters in records, the item is remembering them, seeking them and handling them

speedier as indicated by the requirements of the environment.

II. LITERATURE REVIEW

SandeepKaur, Rekha Bhatia (2016) considered the method for the acknowledgment of online written by hand characters by utilizing the digitizer tablets or composing cushionst takes a shot at the premise of time defer neural system, which is initially prepared by utilizing the penmanship of numerous journalists to perceive characters. The framework is speed and memory productive. Three diverse neural system classifiers for the acknowledgment of complex examples.These classifier incorporates structure versatile self-sorting out classifier, HMM half breed classifier and numerous multilayer perceptron classifier. Along these lines, it is vital to pick the procedures for building up an OCR system according to the prerequisite of utilization .A methodical stream of OCR framework is examined and past work donein this field is overviewed stage by stage.These are characterized into two classifications: non-parametric and parametric technique. By and large, parametric techniques are favoured than non-parametric strategy. In non-parametric technique preparing information is utilized for classification. In the proposed framework, a chain code is utilized for the representation of the characters and grouping is done on the premise of use of a committed processor for the string examination. Basic and Syntactical Method- Basic and linguistic techniques are utilized to perceive the manually written characters or patterns.

NalinBhat, Avinash Kumar Yadav, KrushnaRajbind (2016) created an OCR which will perceive machine printed composed English, Spanish characters. We can do this by enhancing our product for perceiving the printed content record. Finish change of A4 picture record (printed version) into advanced word i.e.OCR is an Optical character acknowledgment and is the mechanical or electronic interpretation of pictures of typewritten content (more often than not caught by a scanner or camera) into machine-editable text. This framework can be utilized by numerous clients .A product which will perceive the characters from disconnected record (in picture format).Here we are creating OCR which will perceive machine printed composed English, Spanish characters. In this unique circumstance, out-of-center obscure is a noteworthy issue: clients have no immediate control over it, and it truly Degrades OCR acknowledgment.OCR is an Optical character acknowledgment and is the mechanical or electronic interpretation of pictures of typewritten content (ordinarily caught by a scanner) into machine-editable content Optical Character Recognition (OCR) is a system that disentangles pictures of typewritten inspected content into machine-editable substance, or pictures of characters into a standard encoding arrangement addressing them in ASCII or Unicode. An OCR structure enables us to sustain a book or a magazine article clearly into an electronic PC record, and adjust the archive using a word processor. Another strategy is displayed for versatile report picture banalization, where the page is considered as an accumulation of sub parts, for example, content, foundation and picture.

M. Antony Robert Raj & Dr. S. Abirami (2014) showed the use of an OCR in Tamil hand written scripts. A statistical technique utilizes quantitative estimations for highlight extraction, while basic methods utilize subjective estimations for highlight extraction. Offline Tamil manually written records acknowledgment still offers many propelling difficulties to researchers/Recognition of Tamil transcribed scripts is confused contrasted with other western dialect scripts. Challenges still wins in the acknowledgment of typical and additionally irregular composition, inclining characters, comparable molded characters, joined characters, bends thus on amid acknowledgment process. The thought behind an OCR is to recognize and investigate a report picture by separating the page into line components, additionally sub-partitioning into words, and afterward into characters. Segmentation of transcribed records is more unpredictable than sort composed reports.Pre-preparing comprises of a couple sorts of sub procedures to clean the report picture and make it fitting to convey the acknowledgment procedure precisely. Noise Removal: Median Filtering, Wiener Filtering method and morphological operations can be done to remove noise.A manually written archive must be checked and changed over into an appropriate arrangement for preparing

Rahul V.Chaugule& Prof. SachinGodse (2015) recommended diverse calculations for recognizing understudies penmanship for enveloping a programmed, quick and solid framework which evaluate the characteristics of specific understudy from hypothesis exam papers.Text Recognition using Invariant Moments: The procedure of optical image acknowledgment is isolated into two phases. Optical Character Recognition Implementation Using Pattern Matching Contour Analysis Algorithm.Essentially to change over no-editable records to editable archive we are using optical character affirmation (OCR system).For this reason there is Optical character acknowledgment (OCR) framework and by utilizing distinctive calculations of OCR framework we can without much of a stretch change over the picture document to content record so we can look and alter everything from that record. Downsides of existing calculation are that it requires bigger preparing time, requires complex impulse, they utilize wide numerical operations and include numerous figuring to reason the wavelets. To recognize the student's handwriting there are many algorithms some of which are listed below-A Stock Pattern Recognition Algorithm.

Najib Ali Mohamed Isheawy and HabibulHasan (2014) stated that the inspiration driving this OCR structure is to get a handle on physically composed English characters as data, process the character, set up the neural framework count, to see the illustration and alter the character to an upgraded interpretation of the information. This work is confined to English characters and numerals as it were. The idea of putting away the substance of paper reports in PC

stockpiling spot and after that perusing and seeking the substance is called document processing. The greater part of the accessible techniques manage the acknowledgment of the Roman script and a portion of the oriental scripts like Kanji, Kana, and soOptical character affirmation implies the branch of programming building that incorporates examining content from paper and making an understanding of the photos into a frame, the PC can control. Progressed OCR frameworks can read message in expansive assortment of textual styles, yet regardless they experience issues with manually written text.objective vector and furthermore a vector which contains the example data, this could be a picture and written by hand informationThe neural framework then attempts to make sense of whether the data arranges an illustration that the neural framework has retained. The Grid infrastructure utilized as a part of the usage of Optical Character Recognition framework can be proficiently used to accelerate the interpretation of picture based reports into organized archives that are right now simple to find, hunt and process.

III. METHODOLOGY

There are various modules in our system such as document processing, system training for handwritten recognition, document recognition, document editing, document searching etc. There are various methods available to satisfy the above given modules like use of tesseract, kohonen neural network(KoNN) or feed-forward back neural system, [3]. To understand the handwritten characters, the use of neural network is the best thing.

A. Algorithm Used:

Preparing without supervision is that the neural system is furnished with preparing sets, which are accumulations of characterized info values. However, the unsupervised neural system is not furnished with expected yields. Unsupervised preparing is generally utilized as a part of a characterization neural system. An arrangement neural system takes input designs, which are displayed to the info neurons. These information examples are then handled, and one single neuron on the yield layer fires. This terminating neuron can be considered as the arrangement of which gathering the neural information design had a place with. Penmanship acknowledgment is a decent utilization of a grouping neural system. The information designs displayed to the Kohonen neural network(KoNN) are the dab picture of the character that was written by hand. We may then have 26 yield neurons, which relate to the 26 letters of the English letter set. The Kohonen neural system ought to arrange the info design into one of the 26 input designs. Amid the preparation procedure the Kohonen neural system in written by hand acknowledgment is given 26 input designs. The system is designed to likewise have 26 yield designs. As the Kohonen neural system is prepared the weights ought to be balanced so that the information examples are characterized into the 26 yield neurons. This procedure brings about a moderately viable strategy for character acknowledgment. Another basic application for unsupervised preparing is information mining. For this situation you have a lot of information, however you don't frequently know precisely what you are searching for. You need the neural system to order this information into a few gatherings. You would prefer not to direct, early, to the neural system which input example ought to be arranged to which gather. As the neural system prepares the info examples will fall into comparable gatherings. This will permit you to see which input examples were in like manner gatherings. Yield from the Kohonen neural network (KoNN) does notcomprise of the yield of a few neurons. At the point when an example is introduced to a Kohonen arrange one of the yield neurons is chosen as a "champ". This "triumphant" neuron is the yield from the Kohonen orchestrate. Routinely these "triumphant" neurons address bundles in the data that is shown to the Kohonen mastermind .For instance, in an OCR program that utilizations 26 yield neurons, the 26 yield neurons delineate information designs into the 26 letters of the Latin letter set. It is likewise essential to comprehend the confinements of the Kohonen neural system. Neuralsystems with just two layers must be connected to directly distinguishable issues.This is the situation with the Kohonen neural system. Kohonen neural systems are utilized on the grounds that they are a moderately straightforward system to develop that can be prepared quickly.There is no need of two different methodologies for handwritten charactarer recognition and retrieving text from image[8].

B. Structure

The Kohonen neural system contains just an information and yield layer of neurons. There is no concealed layer in a Kohonen neural network(KoNN). The contribution to a Kohonen neural system is given to the neural system utilizing the info neurons. These information neurons are each given the gliding point numbers that make up the info example to the system. A Kohonen neural system requires that these sources of info be standardized to the range between - 1 and 1. Showing an info example to the system will bring about a response from the yield neurons. If we had a neural system with five yield neurons we would be given a yield that comprised of five qualities. This is not the situation with the Kohonen neural system. In a Kohonen neural system just a single of the yield neurons really delivers an esteem. Moreover, this single esteem is either valid or false. At the point when the example is introduced to the Kohonen neural system, one single yield neuron is picked as the yield neuron. Along these lines, the yield from the Kohonen neural system is generally the record of the neuron (i.e. Neuron #5) that terminated.

C. How It Learns?

There are a few stages required in this preparation procedure. Generally the procedure for preparing a Kohonen neural system includes venturing through a few ages until the blunder of the Kohonen neural system is underneath satisfactory level. The preparing process for the Kohonen neural system is aggressive. For each preparation set one neuron will "win". This triumphant neuron will have its

weight balanced with the goal that it will respond significantly more emphatically to the info whenever. As various neurons win for various examples, their capacity to perceive that specific example will be expanded. Hebb's algorithm is also used for grid infrastructure concept.

D. Experiment Design

The best way to explain the design is that the image will be scanned, stored and recognized. These images may be stored anywhere in the system or any database. Now there is an option for searching and editing of the images[9]. For handwritten characterrecognition training system is there so that OCR can recognize letters other than English also like indi scripts or devanagri[10]. The Kohonen Neural Network provides this algorithm. Similarly area will be provided for drawing any letter and then the user can train it and after training the system will recognize all the letters of that language. The user can draw two letter and then give training to system after the training the system will recognize all the words written by the user in that language. The database can be handled by the administrator only and the document searching and editing can be done by both administrator as well as user. A GUI is provided where the administrator or the end user will have the option for Handwritten Recognition, Scanned Images, Text Editor Help.

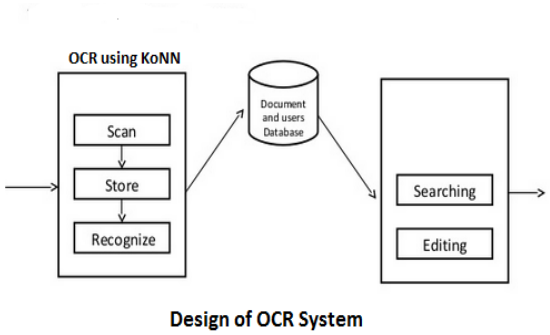

Figure1: Architecture of OCR

As we can see in the figure how the algorithm is used to train the system, at which point it is applied and how we are getting the output. The grid structure is best so as to recognize any language or character. The training data set is given and by the help of kohonen neural network we train the system.

E. *Parameters or Data Used*

As the name suggests that the system is about optical character recognition so we will recognize something. Now let us see what are the data sets or parameters in the system .Firstly, we will talk about handwritten recognition, so for this the data sets are the letters of different languages as our system can recognize letters other than English also,pattern based recognition like signature[14]. So after giving training the system recognizes. Now coming to image part, the data sets cam be documents like newspaper, books etc. If we take an example of a librarian so the data sets will be books or scanned copy of books. Some personal documents can also be scanned and taken electronically. Thesystem can also be used for Autonomous Number Plate Recognition purpose which is of great advantage for security purposes. MRTD (Machine Readable Travel Document) passports can also be one of the parameters[11].The execution of the frameworks have been compelled by the reliance on text style, size and introduction[12].

IV. RESULTS AND DISCUSSION

The accompanying demonstrates thearrangement of yield screens and how the real procedure of executing OCR happens:- The first and the landing page of our optical character acknowledgment framework looks as appeared .It gives an interface to the client with the end goal that the client can get to any module that is available in this product from this page itself. The page is as demonstrated as follows:-

There are two types of recognitions in the document recognition module. They are handwritten letter recognition and the scanned document recognition. The implementation of the handwritten document recognition proceeds as follows:-

We can compose letters on the workspace furnished with the name "Draw Letters Here" by utilizing mouse pointer. For perceiving these letters we need to prepare the framework first. Else, it will give a mistake message portraying that the framework must be prepared first. This procedure is clarified with the accompanying screens:-

Firstly assume that we have drawn a letter named "P" in the workspace given.

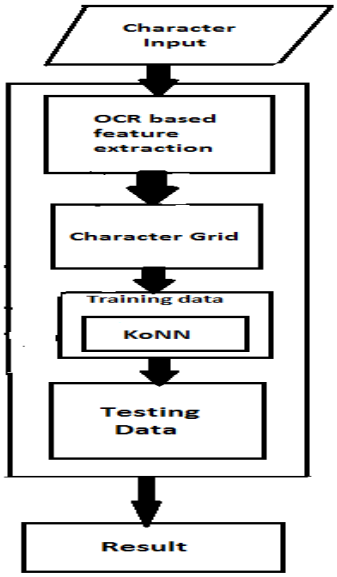

Figure 2: Operational view of OCR using KoNN

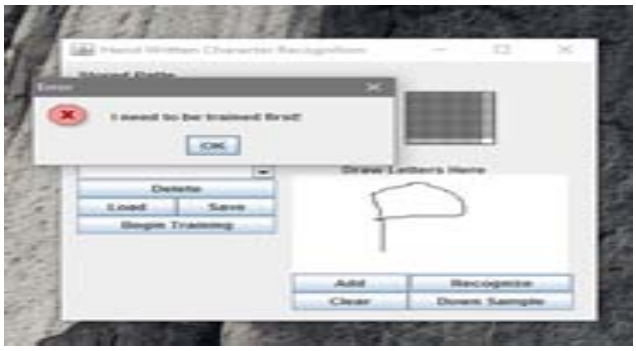

Figure 3: Training of handwritten character

We can click on “Begin training” to train the system and make system to recognize the Character.We can store the pattern also ,like in this image we have stored P and the output says the character is P.

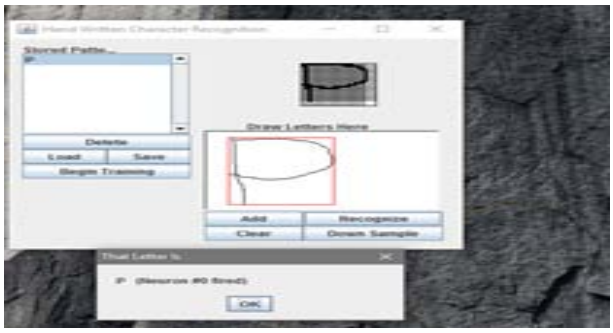

Figure 4: Recognition of character

Examined Document Recognition

There are two principle tabs under the examined archive acknowledgment. They are recognize and training. In the first place we ought to prepare the framework under recognize module. At exactly that point we can perceive the characters from the info picture gave utilizing the training module. The preparation tab under filtered archive training resembles this. At last tap the "recognize catch with the end goal that it extricates/perceives the characters from the picture and shows it to the client. In any case, this information is as yet not editable. Consequently when we tap on the "edit" catch gave at the base focus then therecord gets to be distinctly both editable and searchable.

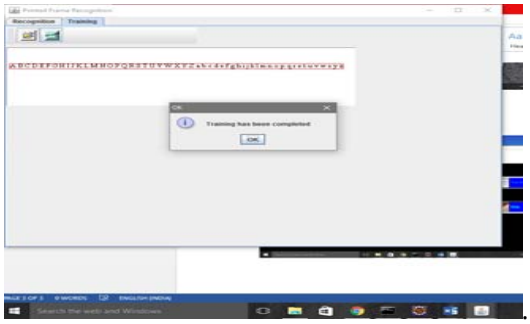

Figure 5 : Training for font

We can roll out any kind of improvements to the archive utilizing cut, duplicate, glue and so forth and you can at last spare the report in two formats (word, content) according to our outline. The hunt capacity can be done here by tapping the "search" picture catch at the base left corner. At that point it requests that the client enter the pursuit.

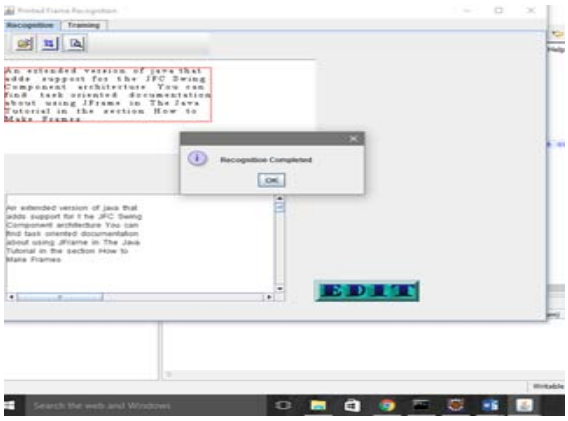

Figure 6: Recognition of Content

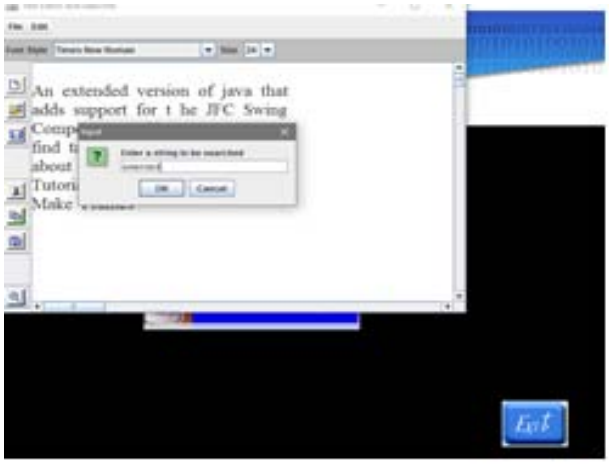

Figure 7: Word Search

If the User is not able to do anything or if the user is facing any issue,he/she can click on help tab and get the issue settled.

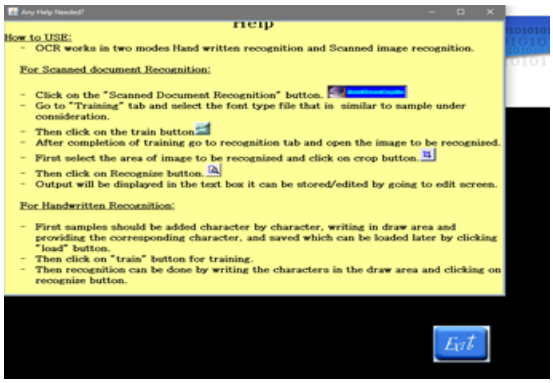

Figure 8: Help

V. CONCLUSION

What does the future hold for OCR? Sufficiently given entrepreneurial planners and adequate innovative work dollars, OCR can turn into an effective instrument for future information passage applications. Be that as it may, the constrained accessibility of assets in a capital-short condition could confine the development of this innovation. In any case, given the correct driving force and consolation,

a considerable measure of advantages can be given by the OCR framework. The computerized section of information by OCR is a standout
amongst the most appealing, work decreasing innovation .The acknowledgment of new text style characters by the framework is simple and fast. We can alter the data of the archives all the more advantageously and we can reuse the altered data as and when required. The augmentation to programming other than altering and hunting is point down future works .The Grid foundation utilized as a part of the execution of Optical Character Recognition framework can be productively used to accelerate the interpretation of picture based records into organized reports that are right now simple to find, pursuit and process. The proposed system can very accurately recognize the handwritten character and printed documents. Like if we draw 'P' it won't recognize so first we will train the system then all the English alphabets will be recognized. Similarly if we draw a Tamil letter using mouse pointer it won't recognize so again we will have to train the system, after the training is completed all the Tamil letters will be recognized by the system. Similarly we can modify the printed document which is either stored in our computer or we can scan and do any kind of modifications or search something in that printed document electronically. The Optical Character Recognition programming can be improved later on in various types of routes, for example, Preparing and acknowledgment rates can be expanded more prominent and more prominent by making it more easy to understand. Numerous applications exist where it is attractive to peruse written by hand passages. Perusing penmanship is an extremely troublesome assignment considering the diversities that exist in customary handwriting. In any case, advance is being made.

V. ACKNOWLEDGMENT

It is incredible chance to expound on subject like Optical Character Recognition .At the season of setting up this paper we have experienced distinctive books and sites which helped us amid this exploration.

We are thankful to our teacher Mr. .Sathyaraj R whose aptitude ,understanding , liberal direction and bolster made it possible for us to chip away at a subject which was extraordinary enthusiasm to us. It was a delight working with him.

VI. REFERENCES

[1].Singh, S. (2013). Optical character recognition techniques: a survey. Journal of emerging Trends in Computing and information Sciences, 4(6), 545-550.

[2].Gupta, V. (2016). Printed Hindi Characters Recognition Using Neural Network.In Proceedings of Fifth International Conference on Soft Computing for Problem Solving (pp. 665-671).Springer Singapore.

[3].Mithe, R., Indalkar, S., &Divekar, N. (2013). Optical character recognition. International Journal of Recent Technology and Engineering, 2(1), 72-75.

[4].Van Deventer, J., Hagen, M., &Mandak, I. (2015). U.S. Patent No. 8,995,774. Washington, DC: U.S. Patent and Trademark Office.

[5].Bhat, N., Yadav, A. K., &Rajbinde, K. (2016). Recognition, Formatting In Image Files By Using Image Processing. Imperial Journal of Interdisciplinary Research, 2(3), 51-54.

[6].Raj, M. A. R., &Abirami, S. (2012). AS urvey on Tamil Handwritten Character Recognition.

[7].Dahake, K. R., Suralkar, S. R., &Ramteke, S. P. (2013). Optical Character Recognition for Marathi Text Newsprint. International Journal of Computer Applications, 62(16).

[8].Zagoris, K., Pratikakis, I., Antonacopoulos, A., Gatos, B., &Papamarkos, N. (2014). Distinction between handwritten and machine-printed text based on the bag of visual words model. Pattern Recognition, 47(3), 1051-1062.

[9].Ghorpade-Aher, J., Gajbhar, S., Sarode, A., Gayake, G., &Daund, P. (2016). Text Retrieval from Natural and Scanned Images. International Journal of Computer Applications, 133(8), 10-12.

[10]..Mathew, M., Singh, A. K., &Jawahar, C. V. (2016, April). Multilingual OCR for Indic Scripts.In Document Analysis Systems (DAS), 2016 12th IAPR Workshop on (pp. 186-191).IEEE.

[11].Do, H. N., Vo, M. T., Vuong, B. Q., Pham, H. T., Nguyen, A. H., &Luong, H. Q. (2016, October). Automatic license plate recognition using mobile device.In Advanced Technologies for Communications (ATC), 2016 International Conference on (pp. 268-271).IEEE.

[12]. Mohammad, F., Anarase, J., Shingote, M., &Ghanwat, P. (2014). Optical character recognition implementation using pattern matching. International Journal of Computer Science and Information Technologies, 5(2), 2088-90.

[13].Mutua, S. M. (2016). An automatic number plate recognition system for car park management (Doctoral dissertation, Strathmore University).

[14].Shalin A. Chopra, Amit A. Ghadge, Omkar A. Parwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar," Optical Character Recognition" International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014

[15]. Tathod, M., Shah, R., Nalamwar, S. R., Yadav, R., & Shah, M. (2013). Devnagari script recognition using kohonen neural network. International Journal on Internet & Distributed Computing Systems, 3(1).