Optical Character Recognition using Artificial Intelligence

Shreshtha Garg
CS Department
Shri Ramswaroop Memorial University, Lucknow

Kapil Kumar Gupta
CS Department
Shri Ramswaroop Memorial University, Lucknow

Nikhil Prabhakar
CS Department
Shri Ramswaroop Memorial University, Lucknow

Amulya Ratan Garg
CS Department
Shri Ramswaroop Memorial University, Lucknow

Aayush Trivedi
CS Department
Shri Ramswaroop Memorial University, Lucknow

ABSTRACT
This is a complete Optical Recognition System using artificial intelligence. In this paper we have dealt with words and character detection. The OCR system that will train itself and help in extracting text from any image by using neural networks and back propagation techniques. Earlier we had to store data in a DATABASE while performing operations but with this we will be able to train our systems. We will have to store only a limited amount of data and the software will self train itself for future entries.

Keywords
OCR, Artificial Intelligence, Neural Network. Feature extraction.

1. INTRODUCTION
Optical Character Recognition, as the name suggests deals with the conversion of pictures that may be handwritten, typed or printed text into machine encoded text electronically on a scanned document, an image of a document with subtitles. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. As far as the document input tasks are concerned, this proves to be a very cost effective and faster method that helps us to free large amount of spaces. Character recognition of printed material is itself a challenging problem since there is a variation of the same character due to type of fonts ,writing format of characters or types of noises[1].It helps in reducing human errors up to a definite limit. As the size and shape of fonts is not the same for all the places it becomes very difficult to recognize the characters. To reduce this problem pre-processing, feature extraction and recognition techniques must be error free. Sometimes the image acquires noise due to electrical distortion, signal change or scanning techniques which distorts the characters and makes it very difficult to extract features. Therefore, a good character recognition approach must eliminate the noise after reading binary picture data, smooth the picture for better recognition, extract features efficiently, train the system and classify patterns.

This paper will tell us that how artificial neural network will help us in increasing the performance of OCR and enable us in recognizing different characters using different datasets and training sets. A neural network is a powerful data processing tool that is able to capture and represent complex input/output relationships. The motivation for the development of neural network technology stemmed from the desire to develop an artificial intelligent system software that could perform "intelligent" tasks similar to those performed by the human brain. Neural networks resemble the human brain in the following two ways: they acquire knowledge through previous learning, and the knowledge is stored within inter-neuron connection strengths known as synaptic weights [1], [2].

Various techniques are used for designing an efficient OCR system. Some of them are Matrix Matching, Fuzzy Logic, Feature Extraction, Structural Analysis, Bayesian Classification and Neural Network. In Matrix Matching, an image is compared to a stored glyph on a pixel-by-pixel basis; it is also known as "pattern matching. This relies on the input glyph being correctly isolated from the rest of the image, and on the stored glyph being in a similar font and at the same scale. This technique works best with typewritten text and does not work well when new fonts are encountered[3]. Fuzzy logic is a multi-valued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, black/white etc [4]. Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations [5]. Structural Analysis identifies characters by examining their sub features- shape of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints [4].And finally the neural network that is explained above.

In this project we will deal with both training and extraction of text from images that would help people in real time and work with huge amount of data easily. By using this application children would be able to learn pronunciations and improve their reading skills as the text gets converted to speech. The specially disabled people would be able to gain information and learn by listening. The process of OCR contains steps including segmentation, feature extraction, and classification. The reason behind the development of the technique is the advent of Unicode and support of complete scripts on computer. The system components proposed includes, 1)Image Scanning, 2)Image Segmentation, 3)Noise Removal 4)Feature Extraction 5)OCR Post Processing and 6)Output from the image in proper Format.

2. PREVIOUS WORKS
Character recognition may be a set of pattern recognition space. to duplicate human functions by machines associate degreed creating the machine perform common tasks like

reading is an ancient dream. The origins of character recognition dates back to 1870 when C.R.Carey of Boston Massachusetts [6, 7, 8] fabricated tissue layer that was a picture transmission system using a mosaic of photocells. The history of OCR are often derived as early as 1900, when the Russian scientist Tyuring trying to develop an aid for the visually handicapped [9]. The first character recognizers appeared in the middle of the 1940s with the development of digital computers [6].

After 20 years Nipkow [15] developed sequential scanner which was a major breakthrough both for modern television and reading machines. During the first few decades of Nineteen century several attempts were made to develop devices to aid the blind through experiments with OCR [8].

By 1950 the technological revolution [6, 9] was moving forward at very fast speed and e-data processing was become an upcoming and important field. The commercial character recognizers available in 1950s where electronic tablets captured the x–y coordinate data of pen tip movement were first introduced. This innovation enabled the researchers to work on the online handwriting recognition problem [6]. The data entry was performed through punched cards. A cost effective way of handling the increasing amount of data was then required. At the same time the technology for machine reading was becoming sufficiently mature for application. By mid 1950s OCR machines became commercially available [8].The first OCR reading machine [16] was installed at Reader's Digest in 1954. This equipment was used to convert typewritten sales reports into punched cards for input into the computer.

In 1966 a thorough study of OCR requirements was completed and an American standard OCR character set was defined as OCR–A. This font was highly stylized and designed to facilitate optical recognition although still readable to humans. An European font was also designed as OCR–B that had more natural fonts than American standard. An Attempt was made to merge two fonts in one standard through machines which could read both standards.

Actual progress on OCR systems achieved during 1990s using the new development tools and methodologies which are empowered by the continuously growing information technologies. In the early 1990s, image processing and pattern recognition techniques were efficiently combined with artificial intelligence methodologies. Researchers developed complex OCR algorithms, which receive high-resolution input data and require extensive number crunching in the implementation phase. Now a days, in addition to the more powerful strong computers and more accurate electronic tools such as scanners, cameras, and electronic tablets, we have efficient, modern use of methodologies such as artificial neural networks(ANNs), hidden Markov models (HMMs), fuzzy set reasoning and natural language processing. The recent systems for the machine printed offline [17, 6] and limited vocabulary, user dependent online handwritten characters [17] are quite satisfactory for restricted applications. However, still a long way to go in order to reach the ultimate goal of machine simulation of fluent human reading especially for unconstrained online and offline handwriting.

Rahul KALA and team proposed work on Offline Handwriting Recognition method with used Genetic Algorithm. In this research describe to a piece of paper and then convert it into text. They are used genetic algorithm to implement Offline Handwriting Recognition [10]. Brandon Maharrey COMP 6600 Artificial Intelligence Spring 2009 they survey about A Neural Network Implementation of Optical Character Recognition that measure that neural network is also use in OCR for the handwritten notes or words [11]. Sang Sung Park, Won Gyo Jung, Young et.al: teams they are implemented Optical Character Recognition System Using BP Algorithm they told her They use OCR (OCR: Optical Character Recognition) technique which is that saving relevant documents to DB after extracting text by using OCR. That is, text should be entered to DB after classifying segments one by one in realized whole document after doing character recognition through OCR. In this paper, in order to solve this problem, we constructed OCR system that saves abstracted characters to DB automatically after extracting only equivalent and necessary characters from a large amount of documents by using BP algorithm [12]. Made Edwin Wira Putra, Iping Supriana Suwardi both has implement the Structural off-line handwriting character recognition using the purpose of those model is to give the ability in improving recognition accuracy without relying in normalization technique. They are use in graph technique. The graph consists of several edges that indicate the connected vertices. The vertices are joining and to form a curve that make the character. The curve is extracted by analyzing the character's chain code, and its string feature is created using some principle [13]. Krupa dholakia has to define about the handwritten character recognition technique are divided into some subparts such as preprocessing, segmentation, feature extraction, classification and post processing[14].

Dr.Mrs.V.V.Patil , Rajharsh Vishnu Sanap , Rohini Babanrao Kharate carried out a study handwritten character recognition using Artificial Neural Network. Artificial neural networks are commonly used to perform character recognition due to their high noise tolerance. The systems have the ability to yield excellent results. The feature extraction step of optical character recognition is the most important. A poorly chosen set of features will yield poor classification rates by any neural network. A simplistic approach for recognition of Optical characters using artificial neural networks has been described[20]. Sanjay Kumar, Narendra Sahu , Aakash Deep , Khushboo Gavel, Miss Rumi Ghosh have introduce the offline handwriting recognition, to provide in improving the ability of recognition accuracy. In this they used the neural network method to implement the OCR. They had researched the Indian doctor prescription and decided to implement the OCR in medical point of view. In this method unknown word (i.e. handwritten word) is input and recognition text is output [19]. Chowdhury Md Mizan, Tridib Chakraborty and Suparna Karmakar present a brief survey of the applications in various fields along with experimentations into few selected fields. The proposed method is very much efficient to extract all kinds of bimodal images including blur and illumination. The paper will act as a good literary survey for researchers starting to work in the field of Optical Character Recognition. [18]

3. PROPOSED WORK

In this we have used Optical Character Recognition with Neural Network. Our work starts with the acquisition of a given image and then scanning the given image for editing and producing desirable outputs. We then pre-processed our images and applied segmentation on it so that the text could be separated. After this the features of the image were extracted which were then followed by recognition and classification.

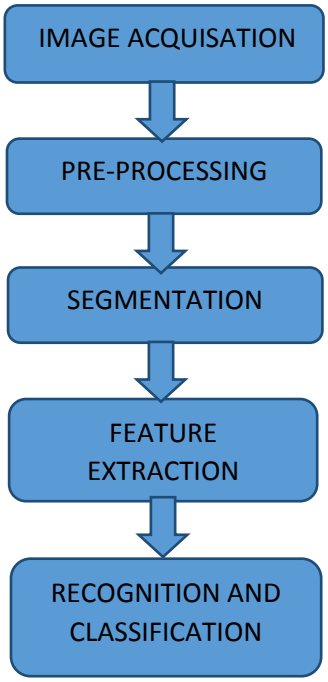

Fig1.Components of OCR with Neural Network

The detailed process is listed below:

3.1. Acquisition of image

In this we read or scan the image and produce a digital image. We do this with the help of an optical scanner that converts light intensity into gray scale.

3.2. Pre-Processing

In this we aim to reduce the color or noise in an image. The scanned image might have some noise which needs to be improved for making the recognition of characters easy and efficient. This is done by applying smoothing on the scanned image and reducing the background noise, like for salt and pepper we use the median filter for smoothing.

3.3. Segmentation

In this process we try to binaries the image for contrasting the text with the background. It is one of the major steps in OCR and with good segmentation we can improve the working of OCR. In this we basically extract the main components of our image i.e. the text that we further use for recognition. It can be of three types – line , word or character segmentation.

3.4. Feature Extraction

In this the important features of the symbols are extracted leaving out the unimportant attributes. We aim to extract the connected pixels. The features can be extracted using different techniques like the distribution of points, transformations and series expansions and structural analysis of characters.

3.5. Recognition and Classification

In this we try to test the features with our existing dataset. If the features match the result is displayed else we update the database with that feature. For classification we use the ANN approach. The data is classified based on the mean distance, spatial position of pixel and pixel value. Here, we use apply training and recognition using the back propagation technique.

4. RESULT ANALYSIS

Table1: Input image size with luminance

SIZE IN BYTES	LUMINANCE
24542	118.94
328580	144.54
479960	149.33
389360	150.57
467910	105.79
905080	134.03
971820	122.67
808690	111.61
319560	175.42
360820	64.238
750670	134.33
290310	147.97
613000	101.68
142640	169.24
1930600	142.59
154900	191.37
92354	157.38
168730	101.17
221630	158.88
117950	39.775
85293	132.3
86947	17.815
78944	123.04
111400	106.72
77248	28.036
39662	216.57
141980	231.75
72648	237.33
33480	171.38
77815	193.75
17383	150.97
17683	208.45
18894	146.41
5873	86.323
43415	180.37
10560	168.92
39662	216.57
337520	140.08

11155	49.519
8918	40.561

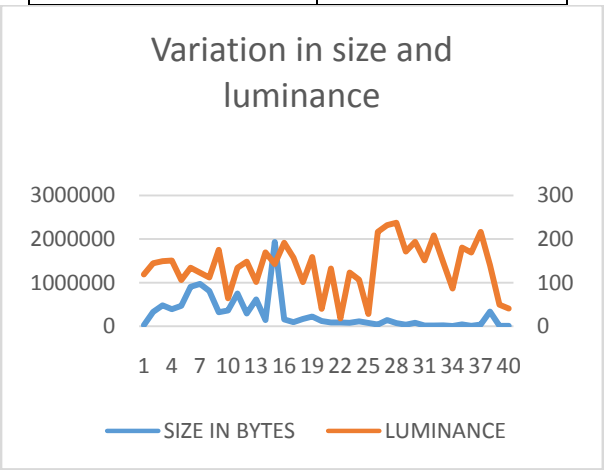

Fig2. Comparison between size of image and luminance

Table2: luminance of images with recognition time

LUMINANCE	TIME IN MILLIS
118.94	74.229
144.54	133.73
149.33	218.47
150.57	865.6
105.79	606.84
134.03	430.09
122.67	914.51
111.61	461.63
175.42	254.42
64.238	224.55
134.33	9833.5
147.97	228.6
101.68	1400.9
169.24	63.856
142.59	3268.5
191.37	207.24
157.38	89.037
101.17	98.947
158.88	471.62
39.775	90.419
132.3	75.694
17.815	30.867
123.04	29.908
106.72	78.36
28.036	18.568
216.57	833.07
231.75	2426.2
237.33	1251
171.38	488.79
193.75	466.11
150.97	232.93
208.45	428.83
146.41	327.36
86.323	71.843
180.37	196.2
168.92	266.31
216.57	831.62
140.08	3278.6
49.519	40.708
40.561	24.9

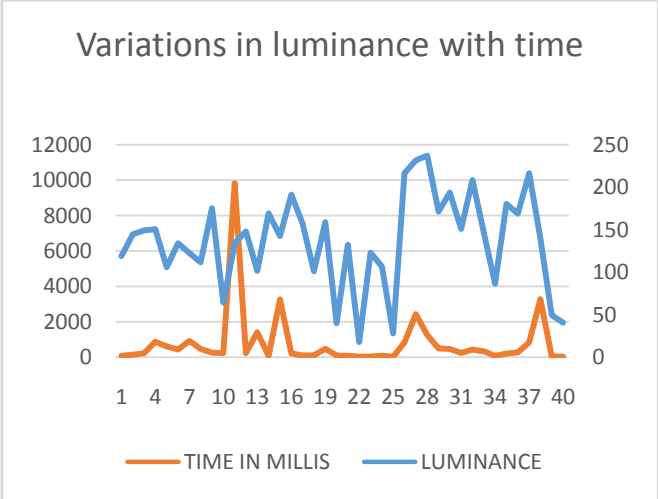

Fig3. Comparison between luminance and character recognition time

Table 3: Size of images with recognition time

SIZE IN BYTES	TIME IN MILLIS
24542	74.229
328580	133.73
479960	218.47
389360	865.6
467910	606.84
905080	430.09
971820	914.51
808690	461.63
319560	254.42
360820	224.55

750670	9833.5
290310	228.6
613000	1400.9
142640	63.856
1930600	3268.5
154900	207.24
92354	89.037
168730	98.947
221630	471.62
117950	90.419
85293	75.694
86947	30.867
78944	29.908
111400	78.36
77248	18.568
39662	833.07
141980	2426.2
72648	1251
33480	488.79
77815	466.11
17383	232.93
17683	428.83
18894	327.36
5873	71.843
43415	196.2
10560	266.31
39662	831.62
337520	3278.6
11155	40.708
8918	24.9

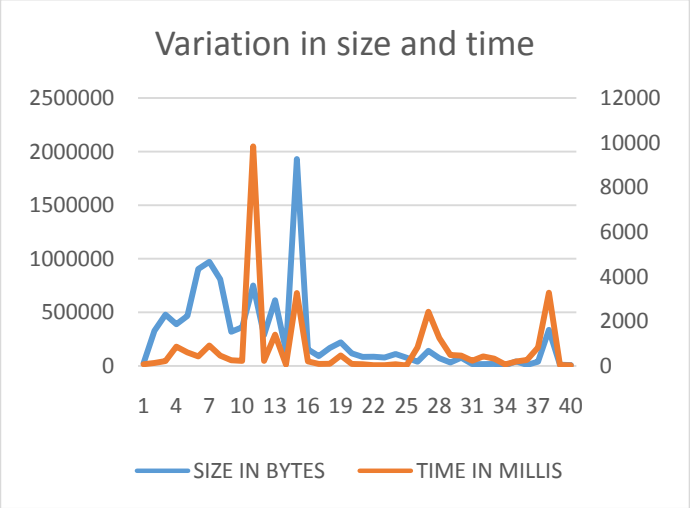

Fig4. Comparison between size of image and character recognition time

4.1 PLATFORM FOR EVALUATION

The platform utilized to evaluate the proposed approach includes a dual core CPU, the Intel Core 2 Duo with clock rate 2.0 GHz and memory 4 GB DDR2 667. MATLAB is used for simulation of code and verifying result.

4.2 TEST SAMPLES OF IMAGES

To validate our proposed approach we tested our assumptions and techniques on various images that were both noisy and noiseless. Some of them are listed below:

Fig.5. Blur image

Fig.6. Noiseless grayscale image

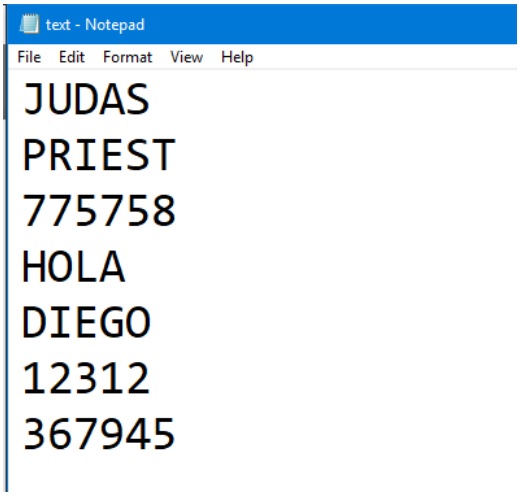

Fig.7. Output received of image in Fig. 6 after passing through OCR with AI

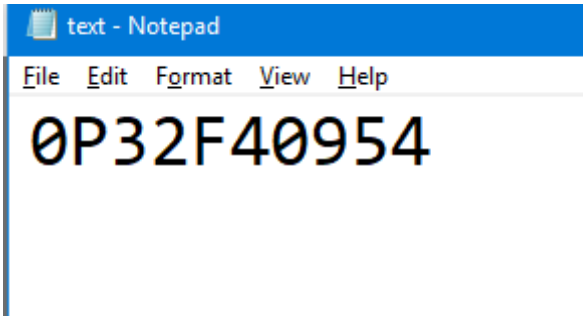

Fig.8 Output received of image in Fig. 5 after passing through OCR with AI

Fig.9. Noisy gray scale image

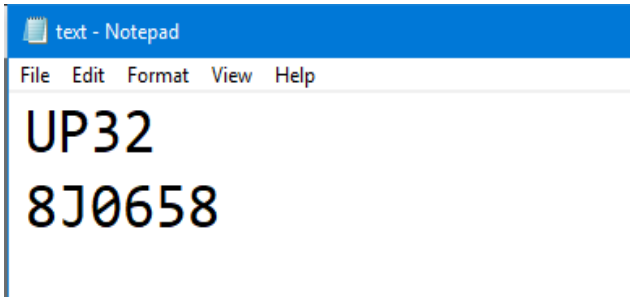

Fig.10 Output received of image in Fig. 9 after passing through OCR with AI

5. CONCLUSION AND FUTURE SCOPE

It was found that the method and approach that was used was able to recognize the character upto 100% when the image was noiseless and almost 95% in noisy images. It also helped in reducing the time for carrying out the whole procedure.

The future enhancements that can be done in this are that it can be trained for hand-written text. For better segmentation and character recognition we can implement the use of dictionary which will help in enhancing the performance of OCR.

6. AKNOWLEDGEMENTS

This work is supported by resources from Dr. Shalini Agarwal, Head of CS Department, Shri Ramswaroop Memorial University, Lucknow and Dr. A.K. Verma Director IOT Shri Ramswaroop Memorial University, Lucknow for promoting research activities in the Campus. We would also like to thank the anonymous reviewers for their significant and constructive critiques and suggestions, which substantially improved the quality of this paper.

7. REFERENCES

[1] Sameeksha Barve "Optical Character Recognition Using Artificial Neural Network" International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 4, June 2012 ISSN: 2278 – 1323.

[2] R.Arnold, Poth Miklos" Character recognition using neural networks", Computational Intelligence and Informatics (CINTI), Hungary, pp 311-314 , 2010.

[3] Herbert F Schantz, "The history of OCR, optical character recognition", [Manchester Center, Vt.]: Recognition Technologies Users Association, 1982.

[4] "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network" Raghuraj Singh, C. S. Yadav, Prabhat Verma, Vibhash YadavChaudhuri, A., Some Experiments on Optical Character Recognition Systems for different Languages using Soft Computing Techniques, Technical Report, Birla Institute of Technology Mesra, Patna Campus, India, 2010.

[5] Isabelle Guyon, André Elisseeff, " An Introduction to Feature Extraction" in Studies in Fuzziness and Soft Computing , Springer pp 1-25 book series (STUDFUZZ, volume 207) 1995-2018.

[6] Rice, S. V., Nagy, G., Nartker, T. A., Optical Character Recognition: An Illustrated Guide to the Frontier, The Springer International Series in Engineering and Computer Science,Springer US, 1999

[7] Schantz, H. F., The History of OCR, Recognition Technology Users

[8] Association, Manchester Centre, VT, 1982.

[9] Mantas, J., An Overview of Character Recognition Methodologies, Pattern Recognition,19(6), pp 425–430, 1986.

[10] Rahul KALA, Harsh VAZIRANI, Anupam SHUKLA,and Ritu TIWARI "Offline Handwriting Recognition using Genetic Algorithm" International Journal of Computer Science Issues, Vol. 7, Issue 2, No 1, March 2010 ISSN (Online): 1694-0784 , ISSN (Print): 1694-0814

[11] Brandon Maharrey "A Neural Network Implementation of Optical Character Recognition" Technical Report Number CSSE10-05 COMP 6600 – Artificial Intelligence Spring 2009 [8].

[12] Sang Sung Park, Won Gyo Jung, Young Geun Shin, Dong-Sik Jang "Optical Character Recognition System Using BP Algorithm" IJCSNS International Journal of Computer Sci 118 ence and Network Security, VOL.8 No.12, December 2008

[13] Made Edwin Wira Putra, Iping Supriana Suwardi "Structural off-line handwriting character recognition using approximate subgraph matching and levenshtein distance" International Conference on Computer Science and Computational Intelligence (ICCSCI 2015 Procedia Computer Science 59 (2015) 340 – 349

[14] Krupa Dholakia "A Survey on Handwritten Character Recognition Techniques For Various Indian Language" International Journal Of Computer Application (0975-8887) Volume 115-No. 1, April 2015

[15] Young, T. Y., Fu, K. S., Handbook of Pattern Recognition and Image Processing, Academic Press, 1986.

[16] Scurmann, J., Reading Machines, Proceedings of International Joint Conference on Pattern Recognition, Munich, pp 1031–1044, 1982.

[17] Arica, N., Vural, F. T. Y., An Overview of Character Recognition focused on Offline Handwriting, IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews, 31(2), pp 216–233, 2001.

[18] Chowdhury Md Mizan, Tridib Chakraborty* and Suparna Karmakar "Text Recognition using Image Processing" International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May – June 2017, ISSN No. 0976-5697.

[19] Sanjay Kumar, Narendra Sahu , Aakash Deep , Khushboo Gavel, Miss Rumi Ghosh "Offline Handwriting Character Recognition (for use of medical purpose)Using Neural Network" International Journal Of Engineering And Computer Science ISSN: 2319-7242,Volume 5 Issue 10 Oct. 2016, Page No. 18612-18615.

[20] Dr.Mrs.V.V.Patil , Rajharsh Vishnu Sanap , Rohini Babanrao Kharate "Optical Character Recognition Using Artificial Neural Network" International Journal of Engineering Research and General Science Volume 3, Issue 1, January-February, 2015 ISSN 2091-2730.

[21] Yusuf Khan, Kapil Kumar Gupta, Namrata Dhanda " Optical Character Recognition using Intro Sort "International Journal of Computer Applications Volume 127–No.1,October 2015, PP 1-4.