Review on Optical Character Recognition of Devanagari Script Using Neural Network

Ms. Smita Ashokrao Bhopi
Department Of Computer Science, IET
MGM's College of CS & IT,
Nanded, Maharashtra (India)
smitabhopi@gmail.com

Mr. Manu Pratap Singh
Department of Computer Science,
Dr. B. R. Ambedkar University
Agra, U.P.(India)
manu_p_singh@hotmail.com

Abstract**—** During the last decades lot of research work has been done in the field of character recognition on various scripts in various languages. In India peoples are used to speak national language Hindi and spoken by more than 500 million people. Many languages in India, such as Hindi, Marathi and Sanskrit has uses Devanagari as its base script .As compared to English character; Indian script (Devanagri) characters are complicated for recognition. Devnagri script is the basis for many Indian script including Hindi, Sanskrit, Marathi, Kashmiri, and so on. In this paper we present a review of research work that has been done in the field of character recognition in Devanagari script in past.

Keywords*: OCR, Preprocessing, Segmentation, Feature Extraction, ANN.*

*****

I. INTRODUCTION

OCR (Optical Character Recognition) is an emerging field of research in Pattern Recognition. In the world more than 300 million people use Devanagari script. Many languages in India, such as Hindi, Marathi and Sanskrit has uses Devanagari as its base script. Devanagari script has basic set of symbols consists of 34 consonants (or vyanjan) and 18 vowels (or swar)[2].Optical Character Recognition for Devanagari is highly complex. There is one difficulty with the Devanagari script is that a word written in Devanagari can only be pronounced in one way, but not all possible pronunciations can be written perfectly because language is partly phonetic in nature. Optical Character Recognition is a process in which scanned page, a printed document or handwritten document is converted in to ASCII character so that computer can recognize it easily. Due to lot of variations in fonts, size of the written characters; there is difficulty in character recognition. So, to remove difficulties in recognition following stages are used.

There are five major stages in the Character Recognition problem.

1) Scanning
2) Preprocessing
3) Segmentation
4) Feature Extraction
5) Classification
6) Post processing

In preprocessing step noise is removed [2]. After the segmentation of characters artificial neural network (ANN) is used to train the extracted character dataset and it will be then used for classification purpose. Use of artificial neural network techniques improves the performance of character recognition. Type Style and Fonts

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times.
Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

II. BLOCK DIAGRAM OF OCR

Any character recognition system goes under following steps, i.e. Image acquisition, Preprocessing, Segmentation, Feature extraction, classification and post processing. Block diagram of general character recognition system is shown in Figure.

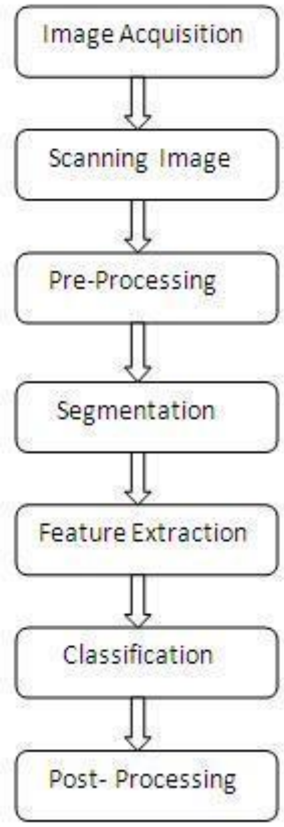


Fig 1. Block diagram of OCR

Following are major stages in the Character Recognition problem.

1) **Scanning:-** First characters written on the hard copy document is get scanned by scanner and then image is converted in to jpg format.
2) **Preprocessing:-**Preprocessing involves following steps
In the proposed OCR system, text is digitized with the help of scanner having resolution between 100 and 600 dpi. The digitized images are usually in gray tone.[1]

Binarization:- This phase consist of the process of converting a gray scale image into binary image by thresholding. Two intensity values are obtained as Black & White.

Size normalization: - As handwritten characters are not uniform in size. So, in order to get characters in uniform size normalization is applied. Each segmented character is normalized in to matrix like 32x32 or 64x64. So, that all characters have same size.

Noise elimination: - In general scanned image may have noise in it. Noise in image is a major problem in character recognition. Due to the presence of noise in the image degrades the quality of image affects on the accuracy in recognition of image. Many filtering techniques are used to remove noise from image. Noise elimination is also called as smoothing. Reduction of noise from the image improves the quality of image. Noise in the image includes distortion, gap in the lines, incomplete corners etc.

Thinning: - The process involves removal of selected foreground pixels from binary image.

3) **Segmentation: -** Segmentation is the process of partitioning an image/document into disjoint and homogenous regions. Segmentation is one of the most important and essential process that that improve the accuracy rate of character recognition system. Devanagari document is partitioned into sequence of lines and words by vertical and horizontal projection respectively. [6].

4) **Feature extraction**:- Feature extraction is defined as extracting the most useful information from the raw data, which minimizes the class pattern variability while enhancing the between class pattern variability [10].Feature extraction is important phase in recognition process and also referred as heart of OCR system. Feature extraction process extracts the most important and relevant shape information present in character.[9].Feature extraction is the special form of Reduction. It reduces the data when input algorithm is very large [8]. Feature extraction methods are broadly classified as Global Transformation and Series Expansion (i.e. Fourier Transforms, Gabor Transform & Wavelets), Statistical Features(i.e. zoning, Projection) and Geometrical and Topological Features (i.e. Extracting and Counting Topological Structures and Coding Graphs & Trees etc) [10]. The methods like histogram of individual characters and GLCM (Gray level co-occurrence matrix) are also considered in feature extraction for character recognition [7].

5) **Classification:** This stage is the decision making step in the optical character recognition system. There are several Classical and soft computing techniques available for handwriting recognition.

Following are the classical techniques used for classification.

a) Template matching: This is one of the basic technique recognition. Matching is used to determine the similarity between two points, curves, or shapes of the same type. In template matching, a 2D shape or a prototype of the pattern to be recognized is available.

b) Statistical techniques: In this statistical approach, each pattern is represented in terms of d features or measurements and is viewed as a point in a d-dimensional space.

c) Syntactic Approach: In this approach, a formal analogy is drawn between the structure of patterns and the syntax of a language. This approach considers patterns are viewed as sentences belonging to a language and primitives are viewed as the alphabet and the sentences are generated according to a grammar.

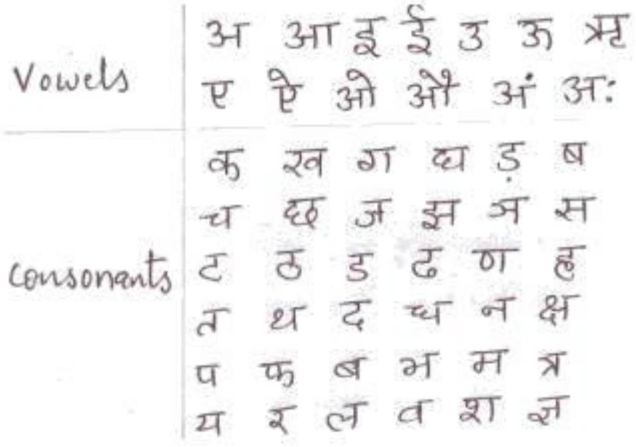
6) **Post Processing:** Post-processing stage is the final stage of the proposed recognition system. It prints the corresponding recognized characters in the structured text form.[14][15].

III INTRODUCTION TO MARATHI SCRIPT

Devanagari word is derived from Sanskrit words Deva (god) and Nagari (city) jointly stand for "city of gods" **[4].**

Devanagari was originally developed to write Sanskrit but was later adopted to write many other languages. Base for every Indian script is Devanagari so called mother of all script. It is used to write languages like Hindi, Marathi, Nepali, Bhojpuri, Bhili, Marwari, Magahi, Maithili etc.[5]

The basic characters of Devanagari script consist of 36 consonants (Vyanjan) and 13 Vowels (Swar). Devanagari script has specific composition rules for joining consonants, vowels and modifiers [6].



Vowels	अ आ इ ई उ ऊ ऋ ए ऐ ओ औ अं अः
Consonants	क ख ग घ ङ ष च छ ज झ ञ स ट ठ ड ढ ण ह त थ द ध न क्ष प फ ब भ म त्र य र ल व श ज्ञ

Fig 2: vowels and 36 consonants of Marathi script.

IV ARTIFICIAL NEURAL NETWORKS

A neural network is a powerful data modeling tool to capture and represent complex input/output relationships. Pattern recognition is extremely difficult to automate. As human brain learn and interact with real world object in the same way Artificial Neural Networks (ANN) develops a computational model that behaves same as human brain interacts and learns new things. ANN consists of a number of units called as Neurons with weighted connections and that work parallel. Learning algorithms are adjusting these weights so as to process information. When this neural network is fully trained we will get the information at the output nodes. The most popular algorithms used are Feed Forward Network, Back Propagation Network, and Radial Basis Function etc. The Back Propagation algorithm determines the weight for a multilayer ANN with feed-forward connections. During the learning phase, the computation is done by minimizing a mean square difference between the desired output and the actual output.

Mathematical Model

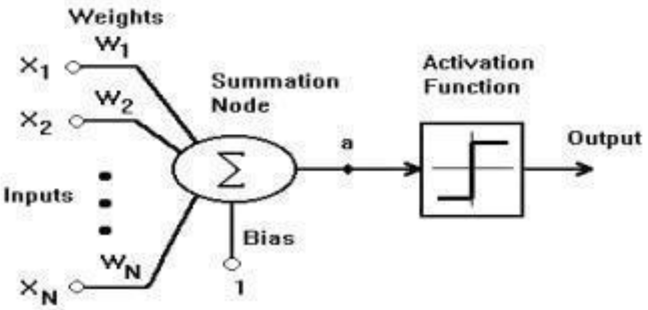


Fig3: Structure of Neural Network

The most common model used in neural network modeling is the multilayer Perceptron (MLP). Supervised learning mode is used in this type of neural network because it requires a desired output in order to learn. This model correctly maps the input to the output. A following fig shows graphical representation of an MLP [11][12].

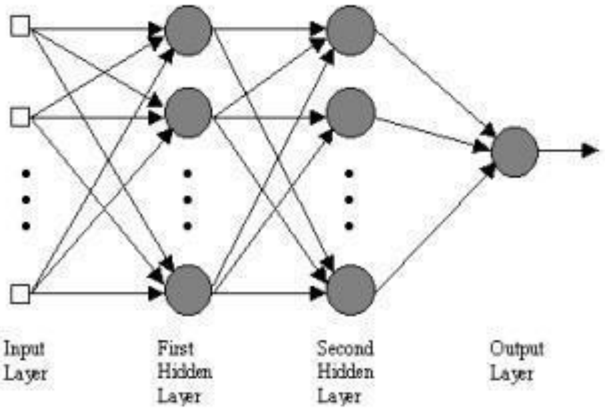


Fig 4:-Block diagram multiplayer Perceptron (MLP).

The inputs are fed into the input layer and get multiplied by interconnection weights as they are passed from the input layer to the first hidden layer. Within the first hidden layer, they get summed then processed by a nonlinear function (usually the hyperbolic tangent). As the processed data leaves the first hidden layer, again it gets multiplied by interconnection weights, then summed and processed by the second hidden layer. Finally the data is multiplied by interconnection weights then processed one last time within the output layer to produce the neural network output.[13]

V RELATED WORK

In this paper authors reviewed methods for pattern representation as statistical methods and classification. Among these two approach author pay more detailed attention towards statistical methods approach. Many neural network algorithms are discussed and various approaches like Template Matching, Statistical Approach, and Syntactic Approach and Neural network for optical character recognition system are discussed.[14]

In the paper authors reviewed and presented a method for Devanagari Optical Character Recognition (DOCR). Author reviewed various methods for character recognition system. Two approaches off- line and on-line character recognition techniques have been discussed. But, as compared to on-line more attention is given to off-line character recognition. Various soft computing methods involved and various classification techniques for optical character recognition like Template Matching, Statistical Techniques, and Neural Networks has been discussed.[16].

Neetu Bhatia in his paper presented a detailed review of Optical Character Recognition and proposed various techniques for character recognition system. Handwriting recognition system divided into types as off-line and On-line character recognition. Where off-line handwriting recognition is difficult and involves automatic of text into an image and on-line involves data stream. Usually OCR involves an off-line character recognition process. Which scan and recognize images of the characters? It is translation of images of handwritten character or into machine code without any variation.[17].

In the paper authors had done a brief survey of Devanagari script and different approach used in classifier for Character Recognition. They proposed method for extraction of feature using local intensity distribution of gradient. They created database of handwritten characters. They have used KNN Classifier. For finding similarity in the pattern they proposed Euclidian Distance method based on K-NN Classification.[18]

In the paper author reviewed the existing works in handwritten character recognition based on Evolutionary computing approach. Author discussed about various approaches like Bio-inspired evolutionary algorithms are probabilistic search methods that simulate the natural biological evolution or the behavior of biological entities, fuzzy approach, and genetic approach.[19]

In the paper authors used the Rectangle Histogram Oriented Gradient representation as the basis for extraction of features.

The algorithm operates on per image pixel. They uses dataset of 8000 samples each of 40 basic handwritten Marathi characters. All sample images are normalized to 20 × 20 pixel size. To obtain result they used support Vector Machines (SVM) and feed-forward Artificial Neural Network (FFANN) classification techniques are used to obtain highest accuracy in Marathi character reorganization.[20]

The author discussed the handwriting recognition systems, evolution and progress. The paper focused on Indic scripts like Bangla, Devnagari, Gurumukhi, Kannada, Malayalam, tamil, and Urdu. The paper focused on multitude of feature and classification techniques and explores new opportunities and challenges in imaging sciences [21].

Author proposes new approaches for extracting features in context of Handwritten Marathi numeral recognition. Artificial Network is used for classification. The overall accuracy of recognition of handwritten Devanagari numerals is 99.67% with SVM classifier, 99% with MLP and it is 98.13with GFF [22]

The paper proposed a new shape based technique for recogntion of isolated handwritten Devnagari characters. Using basic structural features like endpoint, cross point, junction points and adaptive thinning algorithm the thinned character is segmented into segments (strokes). The segments of characters are coded using our Average Compressed Direction Code (ACDC) algorithm. The average accuracy of recognition of the proposed system is 86.4%[23].

The paper presented a new approach for extracting features in context of Handwritten Devanagari Vowels recognition. Artificial Network is used for classification technique. The overall accuracy of recognition of handwritten Devanagari Vowels is % with SVM classifier, % with MLP and it is %with GFF[24].

In the paper, authors discussed the characteristics of the some classification methods that have been successfully applied to handwritten Devnagari character recognition and results of SVM and ANNs classification method, applied on Handwritten Devnagari characters. This process extracted features like shadow features, chain code histogram features, view based features and longest run features. These features are then fed to neural classifier and in support vector machine for classification [25].

This paper guides working on the text based image segmentation area. Author first, the need for segmentation and then, the various factors affecting the segmentation process are discussed. Followed by the levels of text segmentation are explored. Advantages and disadvantages of segmentation are also discussed [26].

This paper focuses on extracting structured data from unstructured data using OCR (Optical Character Recognition) and Neural Network. It focuses on recognizing characters of a document, which does script identification from a variety of unstructured printed or handwritten documents. Discrete Cosine transform function is used to obtain data sets for classification and recognition [27].

The paper describes the behaviors of different Models of Neural Network used in OCR. They mainly focused on parameters like number of Hidden Layer, size of Hidden Layer and epochs. They used Multilayer Feed Forward network with Back propagation. In preprocessing segmentation of characters, normalizing of characters and Deskewing obtained by applying basic algorithms. They have used different Models of Neural Network and applied it on the test data to find the accuracy of the respective Neural Network [28].

This paper presents a Complete OCR system for Marathi text newsprint using Minimum distance classifier [29].

An efficient image retrieval technique is proposed, which uses dominant color and texture features of an image. The attempt is made to enhance the existing results by extracting various supportive features like moments invariant, vector Gradient, chain code (freeman chain code) image thinning, structuring the image in box format, noise removal, etc. A performance of approximately 90% correct recognition is achieved [30].

VI CONCLUSION

Character recognition is one of the important applications of pattern recognition. Day by day popularity of OCR is increasing with the advent of fast computers. But still, lot of research work is needed in OCR to handle the complexity and issues in character recognition.

This paper carries 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 OCR tries to make automation in character recognition to reduce human errors. Artificial neural network is commonly used for training the system. The scanned (input) image is processed and weights are stored and they are used to train data from neural network. Various models like back propagation, multilayer Perceptron used to compare the input image with the trained set to obtain high accuracy in characters recognition.

VII REFERENCES

[1]. “OPTICAL CHARACTER RECOGNITION (OCR) FOR PRINTED DEVNAGARI SCRIPT USING ARTIFICIAL NEURAL NETWORK”, BY RAGHURAJ SINGH1 , C. S. YADAV2 , PRABHAT VERMA3 , VIBHASH YADAV4.

[2]. ”RECOGNITION OF PRINTED AND HANDWRITTEN DEVANAGARI CHARACTERS WITH REGULAR EXPRESSION IN FINITE STATE MODELS”, BY LATESH MALIK, DR. P.S. DESHPANDE PUBLISHED IN DIGITAL TECHNOLOGY JOURNAL 2009, VOL. 2, PP. 1–7, ISSN 1802-5811 (PRINT), ISSN 1802-582X (ONLINE). VSB – TECHNICAL UNIVERSITY OF OSTRAVA, FEECS, 2009.

[3]. ”Devanagari Character Recognition: A Short Review”, by B.Indira ,Muhammad Shuaib Qureshi, Mahaboob Sharief Shaik ,Ramana Murthy ,Rashad Mahmood Saqib in International Journal of Computer Applications (0975 – 8887) Volume 59– No.6, December 2012.

[4]. www.iitm.ac.in

[5]. “PERFORMANCE COMPARISON OF SVM AND ANN FOR HANDWRITTEN DEVANAGARI CHARACTER RECOGNITION”, BY SANDHYA ARORA, DEBOTOSH BHAATTACHARJEE, MITA NASIPURI, L.MALIK, M. KUNDU AND D.K. BASU, INTERNATIONAL JOURNAL OF COMPUTERSCIENCE ISSUES,PP 18-26, VOL. 7, ISSUE 3, NO. 6, MAY 2010.

[6]. “DEVANAGARI CHARACTER RECOGNITION: A SHORT REVIEW”, BY B.INDIRA ,MUHAMMAD SHUAIB QURESHI, MAHABOOB SHARIEF SHAIK ,RAMANA MURTHY ,RASHAD MAHMOOD SAQIB IN INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS (0975 – 8887) VOLUME 59– NO.6, DECEMBER 2012.

[7]. “Optical Character Recognition for Marathi Text Newsprint” by Kiran R.Dahake SSBT’s COET, Bambhori, Jalgaon S.R,Suralkar SSBT’s COET, Bambhori,Jalgaon S.P.Ramteke SSBT’s COET, Bambhori,Jalgaon

[8]. “Character Recognition Using Neural Network” by Ankit Sharma ,Dipti R ChaudharyNirma University Ahemedabad, India. International Journal of Engineering Trends and Technology (IJETT) - Volume4Issue4- April 2013.

[9]. “A REVIEW ON DEVANAGARI OPTICAL CHARACTER RECOGNITION” by Ankush A.Mohod, Prof. Nilesh N.Kasat Sipna College of Engineering & Technology, Amravati. Volume 1, Issue 5, December 2013 International Journal of Research in Advent Technology Available Online at: http://www.ijrat.org

[10]. “A Review of Research on Devnagari Character Recognition” by Vikas J Dongre Vijay H Mankar Department of Electronics & Telecommunication, Government Polytechnic, Nagpur, India. International Journal of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010.

[11]. Handwritten Character Recognition using Neural Network Chirag I Patel, Ripal Patel, Palak Patel in International Journal of Scientific & Engineering Research Volume 2, Issue 3, March-2011 1 ISSN 2229-5518.

[12]. Application of Neural Networks in Character Recognition , V. Kalaichelvi Assistant Professor Dept of Electronics & Instrumentation Engg BITS PILANI, DUBAI CAMPUS Ahammed Shamir Ali Student BITS PILANI, DUBAI

CAMPUS in International Journal of Computer Applications (0975 – 8887) Volume 52– No.12, August 2012.

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

[14]. “Statistical Pattern Recognition: A Review”, by Anil K. Jain, Fellow, IEEE, Robert P.W. Duin, and Jianchang Mao, Senior Member, IEEE.

[15]. “Evolutionary Computing Techniques in Off-Line Handwritten Character Recognition: A Review”, by Gauri Katiyar Shabana Mehfuz. Published in UACEE International Journal of Computer Science and its Applications - Volume 1: Issue 1 [ISSN 2250 - 3765].

[16]. “ A Review of Research on Devnagari Character Recognition”, by Vikas J Dongre Vijay H Mankar ,Department of Electronics & Telecommunication, Government Polytechnic, Nagpur, India. Published in International Journal of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010.

[17]. “Optical Character Recognition Techniques: A Review” , by Er. Neetu Bhatia Kurukshetra institute of Technology & Management Kurukshetra, , India published in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014 ISSN: 2277 128X Research Paper Available online at: www.ijarcsse.com.

[18]. “Brief review of research on Devanagari script”, by Holambe A.N.1 *, Thool R.C.2 , Shinde U.B.3 and Holambe S.N.4. Published in International Journal of Computational Intelligence Techniques, ISSN: 0976–0466, Volume 1, Issue 2, 2010, pp-06-09

[19]. “Evolutionary Computing Techniques in Off-Line Handwritten Character Recognition: A Review “, by Gauri Katiyar Shabana Mehfuz. In UACEE International Journal of Computer Science and its Applications - Volume 1 : Issue 1 [ISSN 2250 - 3765]

[20]. “Handwritten Marathi Character Recognition Using R-HOG Feature”, by Parshuram M.KambleRavinda S.Hegadi in Procedia Computer ScienceVolume 45, 2015, Pages 266-274]

[21]. ”Handwritten Character Recognition: A Review”, by Jayashree Rajesh Prasad, IJCSN International Journal of Computer Science and Network, Volume 3, Issue 5, October 2014 ISSN (Online) : 2277-5420 www.IJCSN.org.

[22]. “Offline Handwritten Devanagari Numeral Recognition Using Artificial Neural Network “, by P E Ajmire. In International Journals of Advanced Research in Computer Science and Software Engineering ISSN: 2277-128X (Volume-7, Issue-8)

[23]. ”Shape Feature and Fuzzy Logic Based Offline Devnagari Handwritten Optical Character Recognition”, by Prachi Mukherji, Priti P. Rege in Journal of Pattern Recognition Research 4 (2009) 52-68 Received Jul 23, 2008. Revised Aug 28, 2008. Accepted Feb 6, 2009.

[24]. ”Handwritten Devanagari Vowel Recognition Using Artificial Neural Network “,by P E Ajmire. In International Journal of Advanced Research in Computer Science, nternational Journal of Advanced Research in Computer Science.

[25]. ”Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition “, by Sandhya Arora1 . Debotosh Bhattacharjee2 , Mita Nasipuri2 , L. Malik4 , M. Kundu2 and D. K. Basu3. In IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, May 2010 www.IJCSI.org.

[26]. ”Segmentation Methods for Hand Written Character Recognition”, by Namrata Dave. In International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 8, No. 4 (2015), pp. 155-164.

[27]. ”OPTICAL CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORK Extracting structured data from unstructured data using OCR and ANN “, by Anamika Bhaduri1 , Deeksha Gulati2 , Sanvar Inamdar3 , Mayuri Kachare 4. In International Journal of Recent Trends in Engineering & Research (IJRTER) Volume 02, Issue 04; April - 2016 [ISSN: 2455-1457].

[28]. ”Handwritten Character Recognition using Neural Network”, by Chirag I Patel, Ripal Patel, Palak Patel. In International Journal of Scientific & Engineering Research Volume 2, Issue 3, March-2011 1 ISSN 2229-5518.

[29]. ”Optical Character Recognition for Marathi Text Newsprint”, by Kiran R.Dahake, S.R,Suralkar, S.P.Ramteke. in International Journal of Computer Applications • January 2013.

[30]. ”Optical character recognition for printed text in Devanagari using ANFIS”, by Prof. Sheetal A. Nirve ,Dr. G. S. Sable in International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013 ISSN 2229-5518.