Handwritten Gujarati text recognition using artificial neural network and Error correction using Probabilistic Neural network in recognized text

[1] Sagar S.Dutt, [2] Prof. Jay D.Amin
[1]Student of Gujarat Technological University,
[2]Assistant Professor at L. J. Institute Of Engineering & Technology ,
[1]Department of Information Technology,
[1]L. J. Institute of Engineering and Technology, Gujarat Technological University, Ahmedabad, Gujarat, India

Abstract

Optical character recognition refers to the process of translating images of hand-written, typewritten, or printed text into a format understood by machines for the purpose of editing, indexing/searching, and a reduction in storage size. Optical character recognition is the mechanical or electronic translation of images of handwritten, typewritten or printed text into machine-editable text. Here we will talk about the Handwritten gujarati character recognition using Artificial neural network and error correction using probabilistic neural network. Character recognition is performed by the Artificial neural networks. Artificial neural networks are commonly used to perform character recognition due to their high noise tolerance. Errors in the character recognition is corrected by the RBPNN(Radial Basis Probabilistic Neural Network) method.

Index Terms— *Optical character recognition, Artificial, Neural Network, Probabilistic neural network Feature extraction, RBPNN (Radial Basis Probabilistic Neural Network)*

I. INTRODUCTION

Optical Character Recognition is one of the most popular applications of image processing and analysis. It is the recognition of printed or written text characters by a computer. This involves photo scanning of the text character-by-character, analysis of the scanned-in image, and then translation of the character image into character codes, such as ASCII, commonly used in data processing.[1]

Documents are scanned using a scanner and are given to the OCR systems which recognizes the characters in the scanned documents and converts them into ASCII data. OCR has three processing steps, Document scanning process, Recognition process and Verifying process. In the document scanning step, a scanner is used to scan the handwritten or printed documents. The quality of the scanned document depends up on the scanner. So, a scanner with high speed and color quality is desirable. The recognizing process includes several complex algorithms and previously loaded templates and dictionary which are crosschecked with the characters in the document and the corresponding machine editable ASCII characters. The verifying is done either randomly or chronologically by human Intervention[2]

II. COMPONENTS OF OCR

The basic steps involved in Optical Character Recognition are:-

1. Image Acquisition
2. Preprocessing
3. Segmentation
4. Feature Extraction
5. Classification and Recognition
6. Post processing

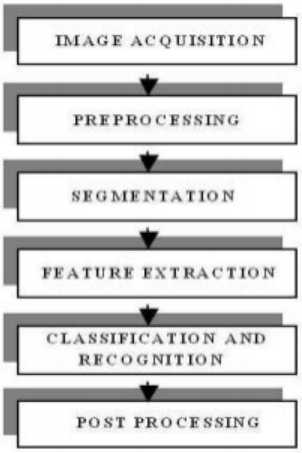

Fig 1. Components of an OCR-system[4]

1] **Image acquisition**

In Image acquisition, the recognition system acquires a scanned image as an input image. The image should have a specific format such as JPEG, BMT etc. This image is acquired through a scanner, digital camera or any other suitable digital input device[4]

2] **Pre Processing**

The image is taken and is converted to gray scale image. The gray scale image is then converted to binary image. This process is called Digitization of image. Practically any scanner is not perfect, the scanned image may have some noise. This noise may be due to some unnecessary details present in the image. So, all the objects having pixel values less than 30 are removed. The noised image thus obtained is saved for further processing. Now, all the templates of the alphabets that are pre-designed are loaded into the system.

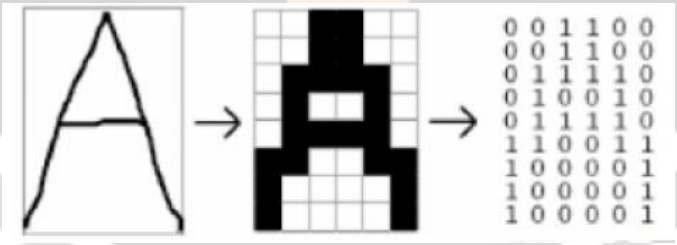

Fig 2. Digitized image[10]

3] **Segmentation**

In segmentation , the position of the object i.e., the character in the image is found out and the size of the image is cropped to that of the template size.

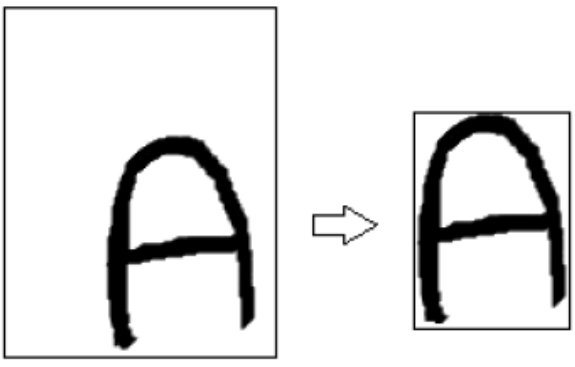

Fig 3. Segmented image[10]

4] **Feature Extraction**

The main objective of feature extraction is to remove redundancy from data. The task of human expert is to select features that allow effective and efficient recognition of pattern. The feature extraction evolves the thinning and skeletonization of the image. Feature extraction is a very important in recognition system because it is used by the classifier to classify the data.

5] **Classification and Recognition**

Classification

The classification is the process of identifying each character and assigning to it the correct character class. In the following sections two different approaches for classification in character recognition are discussed. First decision-theoretic recognition is treated. These methods are used when the description of the character can be numerically represented in a feature vector. We may also have pattern characteristics derived from the physical structure of the character which are not as easily quantified. In these cases the relationship between the characteristics may be of importance when deciding on class membership. For instance, if we know that a character consists of one vertical and one horizontal stroke, it may be either an "L" or a "T", and the relationship between the two strokes is needed to distinguish the characters. A structural approach is then needed.

Recognition

There are two basic types of core OCR algorithm, which may produce a ranked list of candidate characters.[5]

Matrix matching involves comparing an image to a stored glyph on a pixel-by-pixel basis; it is also known as "pattern matching", "pattern recognition", or "image correlation".[6] 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. This is the technique the early physical photocell-based OCR implemented, rather directly.

Feature extraction decomposes glyphs into "features" like lines, closed loops, line direction, and line intersections. These are compared with an abstract vector-like representation of a character, which might reduce to one or more glyph prototypes. General techniques of feature detection in computer vision are applicable to this type of OCR, which is commonly seen in "intelligent" handwriting recognition and indeed most modern OCR software. Nearest neighbour classifiers such as the k-nearest neighbors algorithm are used to compare image features with stored glyph features and choose the nearest match.

Software such as Cuneiform and Tesseract use a two-pass approach to character recognition. The second pass is known as "adaptive recognition" and uses the letter shapes recognized with high confidence on the first pass to recognize better the remaining letters on the second pass. This is advantageous for unusual fonts or low-quality scans where the font is distorted (e.g. blurred or faded).[7]

6] **Post processing**

Recently, In the post processing the use of neural networks to recognize characters (and other types of patterns) has resurfaced. Considering a back propagation network, this network is composed of several layers of interconnected elements. A feature vector enters the network at the input layer. Each element of the layer computes a weighted sum of its input and transforms it into an output by a nonlinear function. During training the weights at each connection are adjusted until a desired output is obtained. A problem of neural networks in OCR may be their limited predictability and generality, while an advantage is their adaptive nature.

III. WHAT IS ARTIFICIAL NEURAL NETWORK?

Artificial Neural Networks are relatively crude electronic models based on the neural structure of the brain. The brain basically learns from experience. It is natural proof that some problems that are beyond the scope of current computers are indeed solvable by small energy efficient packages. This brain modeling also promises a less technical way to develop machine solutions. This new approach to computing also provides a more graceful degradation during system overload than its more traditional counterparts. These biologically inspired methods of computing are thought to be the next major advancement in the computing industry. Even simple animal brains are capable of functions that are currently impossible for computers. Computers do rote things well, like keeping ledgers or performing complex math. But computers have trouble recognizing even simple patterns much less generalizing those patterns of the past into actions of the future. Now, advances in biological research promise an initial understanding of the natural thinking mechanism. This research shows that brains

store information as patterns. Some of these patterns are very complicated and allow us the ability to recognize individual faces from many different angles. This process of storing information as patterns, utilizing those patterns, and then solving problems encompasses a new field in computing. This field, as mentioned before, does not utilize traditional programming but involves the creation of massively parallel networks and the training of those networks to solve specific problems. This field also utilizes words very different from traditional computing, words like behave, react, self-organize, learn, generalize, and forget.

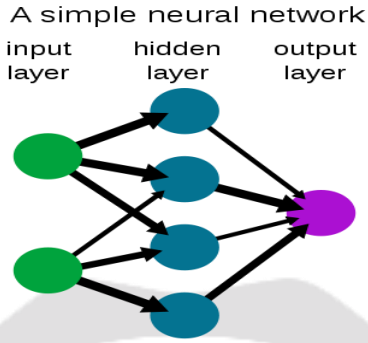

Fig 4.A simple neural network

IV. PROBLEM STATEMENT

Neural Network is Dependant on Character's Font. There should be a different neural network for the different Fonts.

There is always a problem of Degrading paper, wrongly typed and mis-spelled words, ink spreading, missed characters during printing, etc.

Different languages are available. Every language contain various types characters,and every language character contain different properties, so again it's a hard to choose any one technique globally. These things make Printed character recognition is a interesting research field.

Main two parameters are considered by researcher for the good Character recognition system. 1)Accuracy for character Recognition. 2) Time for character Recognition Maintain both parameters is very hard and it's a one of the big challenge with different types of characters.

V. ALGORITHM

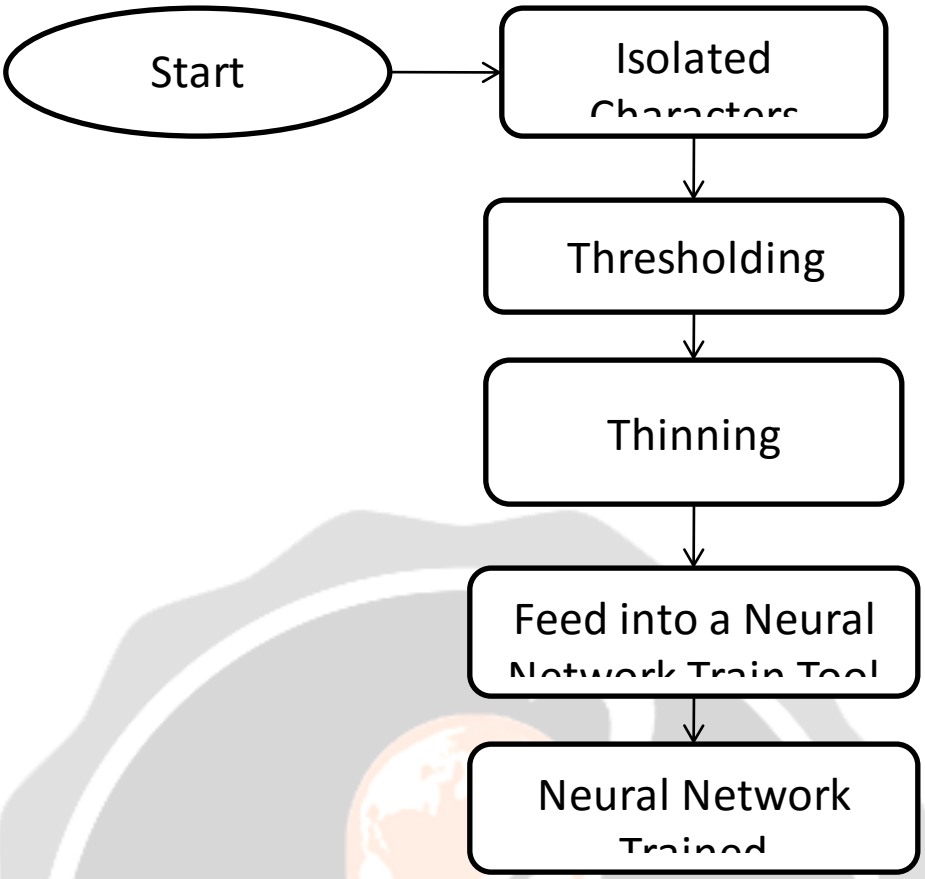

Fig 5.Flow Diagram of proposed system(Training)

Step1:

- Handwritten characters are acquired from the computer in the form of the scanned image of .jpg or .png. The scanned image is divided into the lines, lines are further divided into the words and words are divided into the characters.This is how character isolation would be done.

Step2:

- After that, Thresholding of the character would be done. From a grayscale image, thresholding can be used to create binary images.

Step3:

- After Thresholding the Thinning of the character takes place.It is a morphological operation that is used to remove selected foreground pixels from binary images, somewhat like erosion.

Step4:

- After the Dataset has been prepared, Neural network training is done on that dataset via neural network training tool. There is a dataset of the various handwritten characters. Neural network training is applied on this characters.

Step5:

- After applying the neural network training on the dataset via neural netwok training tool the network is trained.

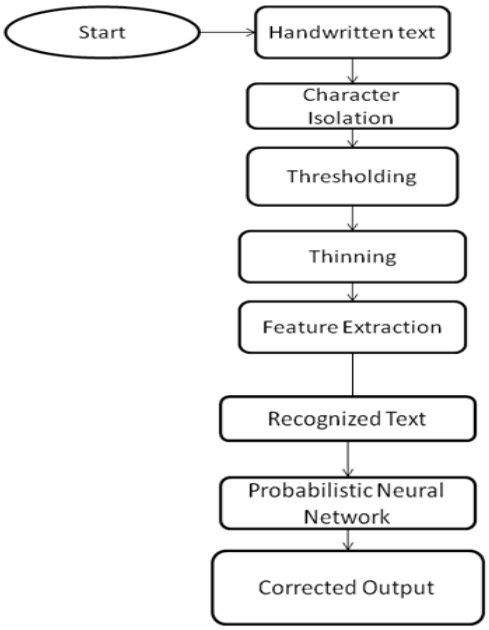

Fig 6.Flow Diagram of proposed system(Testing)

Step1:
Handwritten characters are acquired from the computer in the form of the scanned image of .jpg of .png. The scanned image is divided into the lines, lines are further divided into the words and words are divided into the characters.After character isolation is done.Character thresholding and Thinning process is done

Step2:
After that Thresholding of the character is done.It is the simplest method of image segmentation. From a grayscale image, thresholding can be used to create binary images.

Step3:
After Thresholding, the Thinning of the character takes place.It is a morphological operation that is used to remove selected foreground pixels from binary images, somewhat like erosion or opening.

Step4:
After the Dataset has been prepared, Neural network training is done on that dataset via neural network training tool. There is a dataset of the various handwritten characters. Neural network training is applied on this characters.

Step 5
To train neural network, giving direct images is not a good idea.So we compute some features from the isolated character images in which we have consider sum of rows,sum of columns, sum of left diagonal, sum of right diagonal,height, width and area to match images stored in the database.

Step 6
After applying the feature extraction,the text would be recognized.

Step 7
If there is any error occurred in the recognized text, error correction has to be done.In our proposed method error correction would be done with the help of the Radial basis probabilistic neural network.The training for the dataset would be done with the help of the back propogation algorithm.

VI. IMPLEMENTATION

Step1:

Handwritten characters are acquired from the computer in the form of the scanned image of .jpg of .png. The scanned image is divided into the lines, lines are further divided into the words and words are divided into the characters.After character isolation is done.Character thresholding and Thinning process is done.

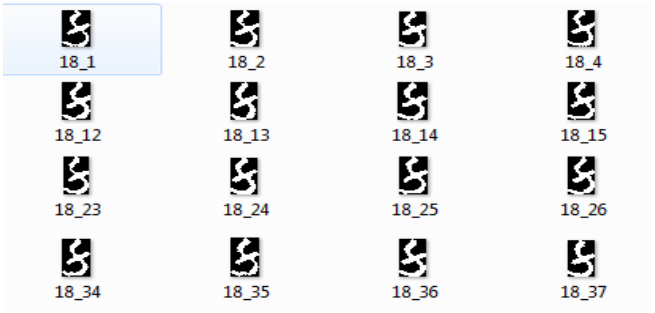

Fig 7.Dataset 0f KA

Step2:

After the Dataset has been prepared, Neural network training is done on that dataset via neural network training tool. There is a dataset of the various handwritten characters. Neural network training is applied on this characters.

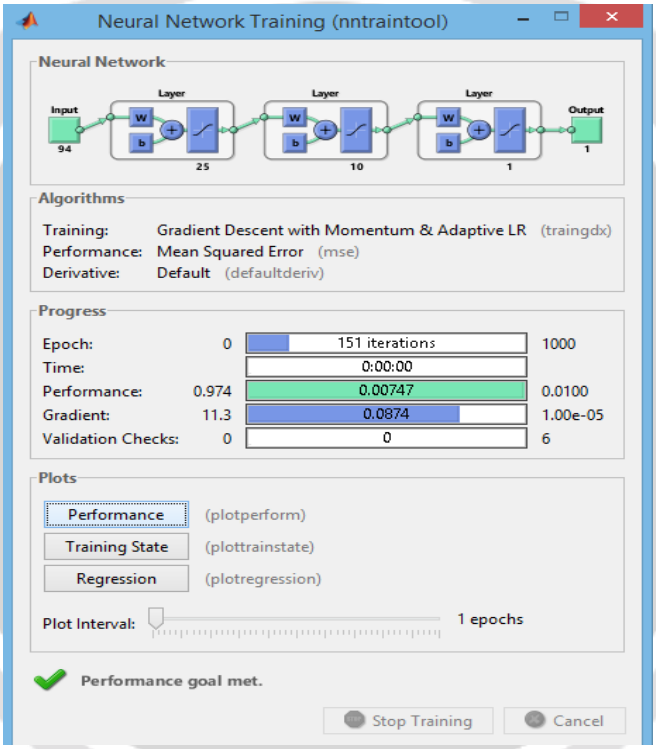

Fig.8 Neural network training tool

Step3:

In step 3 input is given EX. im=imread('0_1.png'); So character of named 0_1 which is "અ" would be recognized from the dataset and it will be shown in the output.

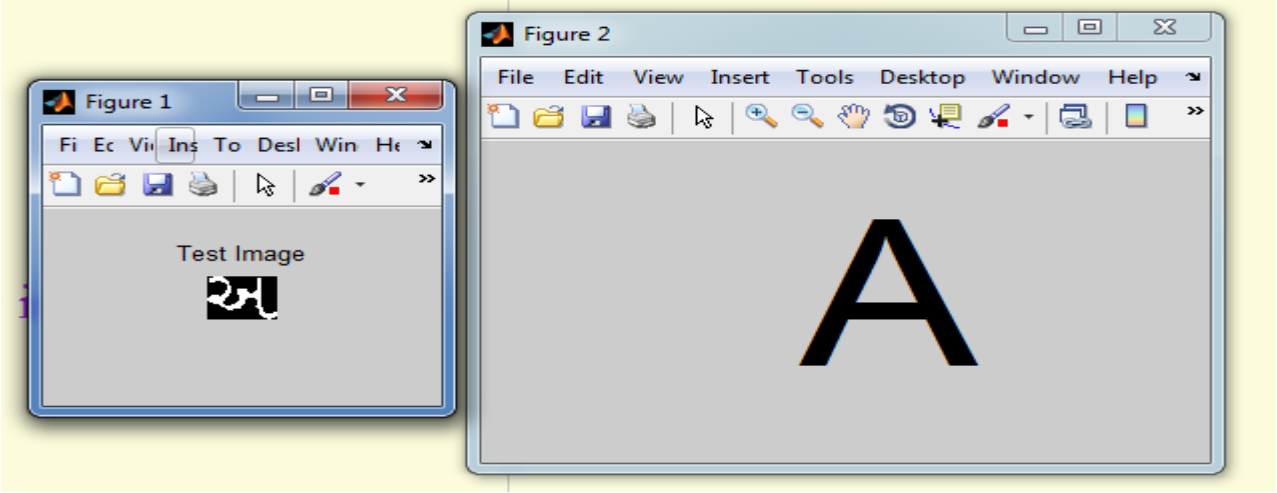

Fig 9 Recognized image ” અ”

Same as step 3 if we changed the input as 1_1(input is given EX. im=imread(‘1_1.png’);) So character of named 1_1 which is “ઘ” would be recognized from the dataset and it will be shown in the output.

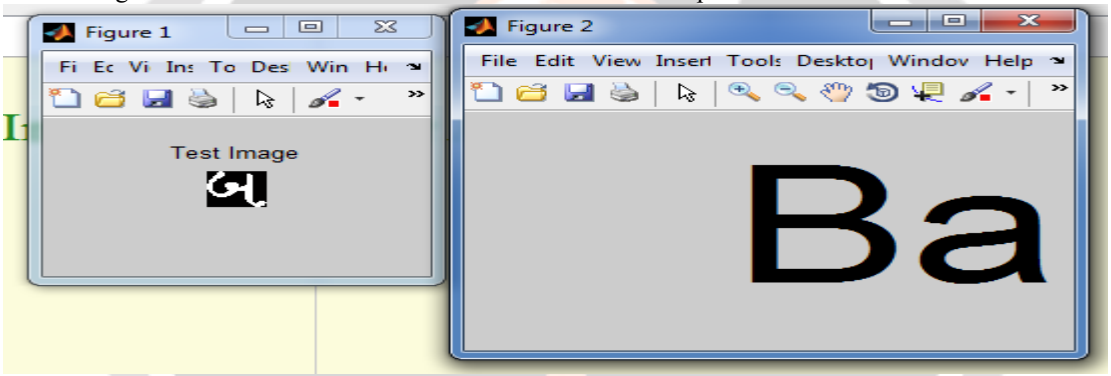

Fig 10 Recognized image ” ઘ”

Results

For ANN Results

(a) Old Approach v/s New Proposed Approach - ANN - Accuracy & Time Parameter - ”trainscg”

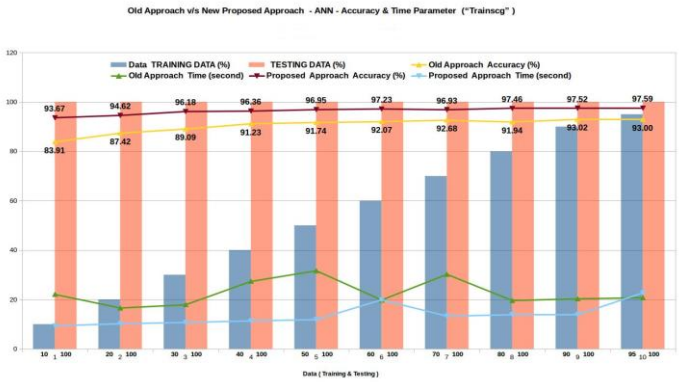

Figure 11: ANN ”trainscg” - Accuracy and Time Parameters

Table 1: ANN Data with ”trainscg” for Accuracy and Time

			Old Approach – Training Methods "Trainscg" 26neuron		Proposed,Approach – Training Methods "Trainscg" 20neuron	
No	Data		Old Approach	Old Approach	Proposed Approach	Proposed Approach
	TRAINING DATA (%)	TESTING DATA (%)	Accuracy (%)	Time (second)	Accuracy (%)	Time (second)
1	10	100	83.905936	22.120587	93.673966	9.41073
2	20	100	87.420229	16.611631	94.62159	10.257125
3	30	100	89.089955	17.938762	96.18389	10.768571
4	40	100	91.231751	27.386296	96.363171	11.409581
5	50	100	91.738788	31.696554	96.952235	11.91438
6	60	100	92.070985	19.860442	97.233961	19.941714
7	70	100	92.682927	30.242183	96.926623	13.326493
8	80	100	91.939855	19.645292	97.464464	13.956719
9	90	100	93.023866	20.322958	97.515687	13.94806
10	95	100	92.99764	20.805109	97.592521	22.776465

Plotted graph shows that proposed approach is gives the batter accuracy and take less time for computation.

(b) Old Approach v/s New Proposed Approach - ANN - Accuracy & Time Parameter - "trainrp"

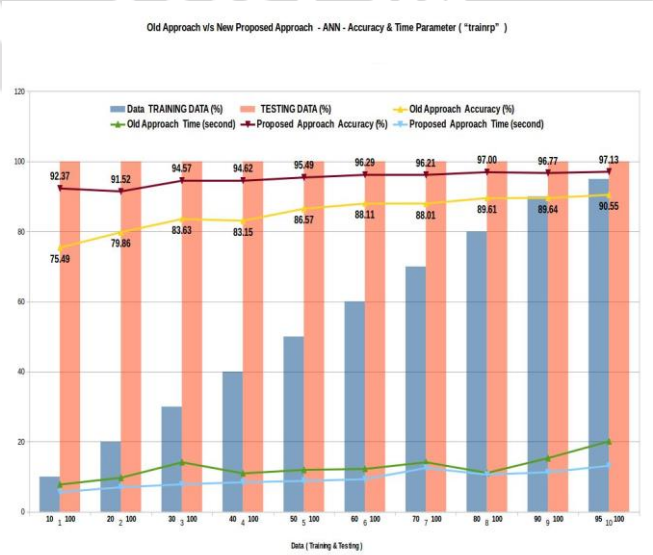

Figure 12: ANN "trainrp" - Accuracy and Time Parameters

Table 2.ANN Data with "trainrp" for Accuracy and Time

No.	Data		Old Approach – Training Methods "Trainrp" 24neuron		Proposed,Approach – Training Methods "Trainrp" 30neuron	
			Old Approach	Old Approach	Proposed Approach	Proposed Approach
	TRAINING DATA (%)	TESTING DATA (%)	Accuracy (%)	Time (second)	Accuracy (%)	Time (second)
1	10	100	75.487368	7.822911	92.367781	5.666271
2	20	100	79.858379	9.771475	91.522602	7.056725
3	30	100	83.634933	14.196153	94.570368	7.89074
4	40	100	83.154122	11.012599	94.62159	8.48964
5	50	100	86.572253	12.02279	95.492381	8.852825
6	60	100	88.110849	12.278405	96.286336	9.357872
7	70	100	88.014687	14.198293	96.209502	12.560835
8	80	100	89.605735	11.150013	97.003458	10.664223
9	90	100	89.640703	15.385084	96.772954	11.354139
10	95	100	90.549873	20.160924	97.131515	13.188602

VII.CONCLUSION

OCR technology provides fast, automated data capture which can save considerable time and labour costs of organisations.Artificial neural networks are commonly used to perform character recognition due to their high noise tolerance. The feature extraction step of optical character recognition is the most important.In this paper there are different Feature extraction techniques and different classification techniques are categorised.

REFERENCES

[1] Jagruti Chandarana, Mayank Kapadia" Optical Character Recognition",ijetae, (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)
[2] Pranob K Charles,V.Harish,M.Swathi, CH. Deepthi" A Review on the Various Techniques used for Optical Character Recognition", (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 1,Jan-Feb 2012,
[3] https://www.nr.no/~eikvil/OCR.pdf
[4] J.Pradeep, E.Srinivasan, S.Himavathi," Diagonal Feature Extraction Based Handwritten Character System Using Neural Network", International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October 2010
[5]"OCR Introduction". Dataid.com. Retrieved 2013-06-16.
[6]How does OCR document scanning work?". Explain that Stuff. 2012-01-30. Retrieved 2013-06-16.
[7]Ray Smith (2007). "An Overview of the Tesseract OCR Engine" (PDF). Retrieved 2013-05-23.
[8]. John, R., Raju, G., Guru, D.S.: 1D wavelet transform of projection profiles for isolated handwritten character recognition. In: Proceedings of ICCIMA07, pp. 481–485, Sivakasi (2007)
[9]. Raju, G.: Wavelet transform and projection profiles in handwritten character recognition—a performance analysis. In: Proceedings of 16th International Co
[10] Sameeksha Barve,"Optical character recognition using artificial neural network", (IJATER), VOLUME 2, ISSUE 2, MAY 2012
[11] Namita Dwivedi, Kamal Srivastava, Neelam Arya," SANSKRIT WORD RECOGNITION USING PREWITT'S OPERATOR AND SUPPORT VECTOR CLASSIFICATION" (ICECCN 2013)

[12] Pritpal Singh, Sumit Budhiraja," Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script ", (IJERA) ISSN: 2248-9622 Vol. 1, Issue 4, pp. 1736-1739

[13] Mayuri Rastogi, Sarita Chaudhary, Shiwani Agarwal," Different Classification Techniques for Character Recognition", MIT International Journal of Computer Science & Information Technology Vol. 3, No. 1, Jan. 2013, pp. 30–34 ISSN 2230-7621

[14] Samit Kumar Pradhan, Sujoy Sarkar and Suresh Kumar Das" A Character Recognition Approach using Freeman Chain Code and Approximate String Matching "published by International Journal of Computer Applications ,2013.

[15] Sameeksha Barve," Artificial Neural Network Based On Optical Character Recognition", (IJERT), ISSN: 2278-0181

[16]Priyanto Hidayatullah, Nurjannah Syakrani, Ida Suhartini, Wildan Muhlis" Optical Character Recognition Improvement for License Plate Recognition in Indonesia" 978-0-7695-4926-2/12 $26.00 © 2012 IEEEDOI 1 0.1109/EMS.2012