A Study Of Devnagri Handwritten Character Recognition System

N. B. Mapari
ME (Final Year, Digital Electronics)
S.S.G.M.C.E, Shegaon
S.G.B. Amravati University (M.S.)
Cell: 9049260200
nagesh_map@rediffmail.com

A. L. Telang
ME (Final Year, Digital Electronics)
S.S.G.M.C.E, Shegaon
S.G.B. Amravati University (M.S.)
Cell: 9850380484
aniltelang_khm@rediffmail.com

R. K. Rajbhure
M.Tech (Final Year, I T)
RGPV University
PCST,Indore.
Cell: 9552320300
ravi.rajbhure@hotmail.com

ABSTRACT

Handwritten characters differ from person to person. Thus, when using traditional methods like Pattern recognition and Image processing techniques, extensive training of the system is needed. Due to this reason, an attempt was made to develop a system that used the methods that humans use to perceive handwritten characters. Thus, a system that recognizes handwritten characters using Pattern recognition was developed.

The reasons for the selection of Pattern recognition were as follows:

Pattern recognition can be used to model human perception.

The mathematics that Pattern recognition requires is extremely fundamental. Thus, any algorithm developed using Pattern recognition would require relatively simple and short calculations.

Due to simplicity of calculations, they can be implemented on any hardware or software platform without too much concern for computing power.

In this paper first part is about introduction to HCR. Then next part giving short introduction for image processing using MATLAB, the actual HCR description the performance, application, and future scope of HCR. Finally last includes the results and conclusion of the paper.

Key terms:

Pattern recognition, image processing, handwritten character recognition, Euclidean distance, nearest neighbor algorithm, database, HCR, MATLAB.

1. OBJECTIVE

The main objective of handwritten character recognition is to interpret the contents of the data and to generate a description of that interpretation in the desired format.

The objectives in the development of this method were, the recognition algorithm used here is very efficient.

The development of a short and efficient algorithm that tries as much as possible to model human perception.

The development of an algorithm that can be implemented on any hardware or software platform through low computational power requirements.

The goal of handwritten recognition is to interpret the contents of the data and to generate description of that interpretation in the desired format with greater accuracy.

2. OVERVIEW

There are three primary processes utilized in most character recognition systems as shown in fig.1. The first is the representation process where giving the input as a character is to get an image of the character and then treated in different ways to achieve a higher level form of the data.

First, the image should undergo some image enhancements such as cropping, reshaping, and filtering out noise, this is called image pre-processing. The raw digitized data is then mapped to a higher level by extracting special characteristics and patterns of the image. This is called feature extraction. The higher level image is then stored in some special way, perhaps in a vector.

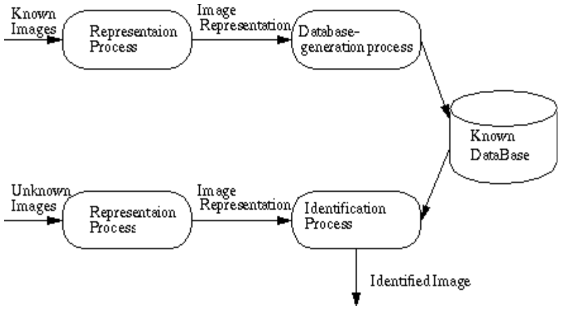


Fig.1: Overview of handwritten character recognition

The second process is the generation of a known database containing the high level representation of all known characters (done in the same way as in the representation process). This is the learning step where the system learns to separate the classes. The database is for knowing how all the different characters could be written and should contain quite a lot of different characters.

In the third part, the identification/classification process classifies the unknown character given its high level representation and the information from the known database. There are several different approaches to make the identification. Here we use statistical method. The statistical approach is based on a similarity measure that is expressed in terms of a distance measure or a discriminate function. Finally a picture of all the processes may perhaps help in understanding the whole idea.

3. INTRODUCTION

The HCR systems have readily been used in a variety of situations and have been highly successful in converting typewritten or handwritten character text into a computer readable format. While HCR is concerned with the full range of alphanumeric characters. In this case, the software tools and methods required to perform recognition are for the constrained version that is for single character.

There are two distinct areas of research concerning Handwritten Character Recognition,

(1) Off-line Character recognition, and

(2) On-line Character Recognition.

Here we use Off-line method. The Off-line character description data provides much more information for recognition. This is due to the reason that these images are usually of characters that were written earlier and later converted to digital format using a digital scanner. The challenge with Off-line character recognition is the development of a system that can recognize these characters. This requires a system that requires very simple and short calculations. If not, the time taken to recognize the characters will render the system useless. Thus, the method selected for this was Pattern recognition.

In this, we present the handwritten character as part from common application Character and of recognizing the alphanumeric characters. The feature extraction algorithm applied to the characters is novel and leads to a very fast recognition.

Although a number of good recognition algorithms have been proposed for handwritten character recognition, the achieved performance is still far from those of human beings in context free handwritten character recognition. The major obstacles have been the different handwriting styles and changeable writing conditions. These two aspects make handwritten characters extremely variable. Observing that the different writing styles or individual writing style of each writer is very important problem in handwritten recognition.

The handwritten identification requires the use of optical mouse as a scanner application for off-line conversion. It is the examination of the design, shape and structure of handwritten character. In this handwriting identification, writer are required to write the same fixed text or also called text- dependent. In practice the requirement for the use of fixed text takes writer identification prone to forgery.

In this technique to determine whether the character wrote is match with the character of the database was presented. This technique is based on computing the main Euclidean distances between these two characters. Then the distance transformation is used to classify whether the handwriting belongs to the same or different character based on the main distances.

4. BASIC DHCR SYSTEM

The basic HCR system is given below in fig.2. This HCR system is divided in two parts. One is training mode and other is recognition mode. During training mode database is prepared and store the results. And during recognition mode sample character is compared with stored patterns in database and computes the result. The input data is given to the system by mouse by drawing character in Paint window or by other form, but must be in digital form.

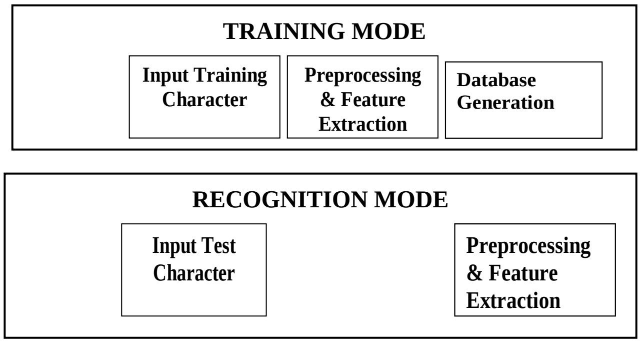


Fig.2: Basic HCR system.

The overall analysis can be divided into four stages: image acquisition for database preparation and preprocessing, feature extraction, and identifying character that is recognition. In all writers were asked to write training alphabets with variation for each alphabet and some testing character through the application interface. The input image is captured from the graphic tablet. At this stage, our system converts image into a 0-1 binary image. A binary image composes of a collection of pixels that are 0 for the background or 1 for the foreground, arranged in a two dimensional matrix.

5. INTRODUCTION TO IMAGE PROCESSING IN MATLAB

5.1 Introduction

This is an introduction on how to handle images in MATLAB. When working with images in MATLAB, there are many things to keep in mind such as loading an image, using the right format, saving the data as different data types, how to display an image, conversion between different image formats, etc. This thesis presents some of the commands designed for these operations. Most of these commands require you to have the Image processing toolbox installed with MATLAB.

For further reference on image handling in MATLAB you are recommended to use MATLAB's help browser. There is an extensive on-line manual for the Image processing toolbox that you can access through MATLAB's help browser.

5.2 Image formats supported by MATLAB:

The following images formats are supported by MATLAB are BMP, HDF, JPEG, PCX, TIFF and XWB. Most images you find in this are BMP images, which are the name for one of the most widely, used compression standards for images. If you have stored an image you can usually see from the suffix what format it is stored in. For example, an image named myimage.bmp is stored in the BMP format and we will use here only BMP images.

5.2.1 Intensity image format:

This is the equivalent to a "gray scale image". It represents an image as a matrix where every element has a value corresponding to how bright/dark the pixel at the corresponding position should be colored. There are two ways to represent the number that represents the brightness of the pixel: The double class (or data type). This assigns a floating number ("a number with decimals") between 0 and 1 to each pixel. The value 0 corresponds to black and the value 1 corresponds to white. The other class is called uint8, which assigns an integer between 0 and 255 to represent the brightness of a pixel. The value 0 corresponds to black and 255 to white. The class uint8 only requires roughly 1/8 of the storage compared to the class double. On the other hand, many mathematical functions can only be applied to the double class.

5.2.2 Binary image format:

This image format also stores an image as a matrix but can only color a pixel black or white (and nothing in between). It assigns a 0 for black and a 1 for white and this is the image we will mostly work with in this.

6. DESIGN OF HCR SYSTEM

The main stages in hcr system are as given below.

1. Database preparation.
2. Preprocessing.
3. Feature extraction.
4. Recognition.

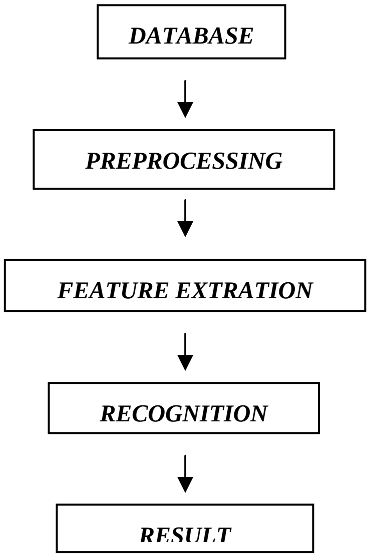


Fig.3: The design flow for HCR.

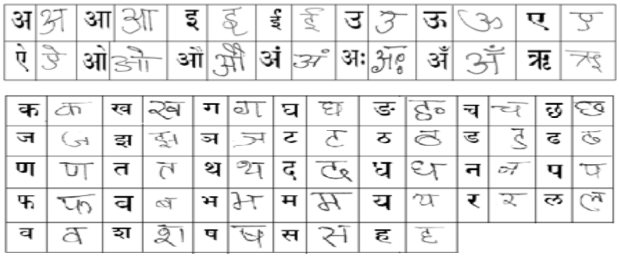
These stages are given in the HCR design flow as shown in fig.3.

6.1 The Database

The entire system is centralized around a main database that contains information about each character, the type of character and the individual features characteristics of the character. The database consists of standard Devnagari handwritten character with different styles of handwriting. Here the database can contain many characters, but an input character can belong to one and only one character from the database. During database preparation, we have to consider lot of things like standard image size, color, format of image. If this parameter does not satisfy, then it fails to recognize. So we have to set this parameter for practical work.

For constructing database of all character, first scan all the character or down load the standard characters from the standard computer library with the use of internet. Here we take characters from the computer keyboard as character image on Paint window, which are store in database as standard .bmp image of the character. The next step is preprocessing of the character images like making the binary images of all data of standard size. Then apply the feature extraction method. Finally load all the respected feature vector of all the character for the comparisons.

The following table 2 shows the database consists of Devnagari handwritten character with different styles of handwriting.



अ	अ	आ	आ	इ	इ	ई	ई	उ	उ	ऊ	ऊ	ए	ए
ऐ	ऐ	ओ	ओ	औ	औ	अं	अं	अः	अः	अँ	अँ	ऋ	ऋ

क	क	ख	ख	ग	ग	घ	घ	ङ	ङ	च	च	छ	छ
ज	ज	झ	झ	ञ	ञ	ट	ट	ठ	ठ	ड	ड	ढ	ढ
ण	ण	त	त	थ	थ	द	द	ध	ध	न	न	प	प
फ	फ	ब	ब	भ	भ	म	म	य	य	र	र	ल	ल
व	व	श	श	ष	ष	स	स	ह	ह				

Table 4: Database of standard Devnagari handwritten character with different styles of handwriting.

6.2 Preprocessing

In the preprocessing, the steps are to digitize the character (scanning). Then apply the image crop operation, normalize it in standard size. Convert digitize character image into the grey scale image and make the binary image of that character from grey scale image. Then apply all the remaining steps of preprocessing listed below

The steps involved in preprocessing are below,

6.2.1 Calculating Threshold:

The first step would be to adaptively threshold the gray level image into the binary image. This was done by calculating the mean and standard deviation of the image and choosing a value equal to their sum as threshold. Then the threshold had to be normalized by dividing by 255. The assumption I have made here is that all the images, which are going to be presented, will be gray level images only.

6.2.2 Conversion of grayscale image into binary image:

After calculating threshold value the next step is to convert the image from gray scale into binary. This was done using standard thresholding procedure. All pixels with intensity values above the threshold would be white and all pixels with intensity values below the threshold would be converted into black.

6.2.3 *Converting to double format:*

In MATLAB we cannot perform some numerical procedures on binary image pixel values. This is because they are stored in the format of unsigned integers. In order to facilitate easy computation the number format of each pixel in the object is converted to double for convenience.

6.2.4 Image cropping:

The aim of dimension reduction (or data reduction) in this case is to throw away unimportant parts of the data, hopefully including the noise, while retaining the information that is important for applying character recognition. As already described, a common image size for a single character can be quite large and result in a very large vector representation. Much less information is required to recognize a character so it is appealing to reduce the dimensionality leaving only those dimensions that are necessary to do recognition. These are often called feature representations because what remains in essence are the important features of the image. The blank space around the edges and corners provide little information, but there is a way of knowing more accurately which dimensions are required. Image cropping is a well known linear dimension reduction technique often used in machine learning and also in variety of statistical applications. This technique is linear in that it does not analyse interactions between different dimensions, which are inherent in the character image, but several non-linear techniques also exist.

The image-cropping algorithm assumes that the data lies close to a hyper-plane and thus vectors that span the hyper-plane alone can represent the data points. The aim is to represent the high dimensional data with a low dimensional representation that will approximate the data and provide the minimum error. The white pixels around the black pixels, which are not useful, are removed. This process is well known as image cropping. Then we resize characters image to some standard pixels window to avoid position of character problem.

6.3 Feature Extraction:

Up until now, the transformation was lossless. As a prerequisite to extracting the features, in order to make the features extracted independent of the input noise. Hence up until this step all (or at least most) of the information in our signal has been preserved.

A feature point is a point of human interest in an image, a place where something happens. It could be an intersection between two lines, any type of distance between pixels or it could be a corner, or it could be just a dot surrounded by space. Such points serve to help define the relationship between different character images. Selecting features according to some criterion amounts projecting it onto a particular view, which usually greatly simplifies the data. A good property of a feature is comparability. That is feature extraction should enable comparison between characters by simple comparisons on the features. In general we will construct a model for each character. Then when a new sample is to be identified, we will see how each model explains the set of features in the sample. The model that explains the combination of features best will be chosen as the most likely category that the sample belongs to.

The recognizing character depends

(1) On the presence or absence of a similar character or nearly the same, and

(2) On the smoothness of this i/p character.

In this, for feature extraction we use strel and imerode functions for getting enlarge view of pixels. This is an important for the viewing feature pixels in the visible form. That simplifies the identification of feature of character on the figure window screen.

Fig. 5 : Printed and handwritten character along with their feature image.

6.4 The recognition mode:

In recognition mode, the user is required to write the character to be recognized on the writing window or tablet that is on Paint platform. The preprocessing and feature extraction steps performed under the training mode for database preparation are then performed as well on the written i/p character. The calculated numerical values of feature vector, which describe the characteristics of the character, are then stored in a variable, thereafter refereed as character feature vector variable. This variable containing information about the character. Once these steps had been performed, focus shifts onto the comparison of the Characters. Then Euclidean distance is calculated between feature vector of both database character and test character. As the values are obtained according to the euclidean distance number of each database character. Then the character is recognized using a min-max Euclidean distance inference over the database characters and test character.

For each characteristic of the characters map each numerical values calculated for the character to be recognized against the information stored in the database for that characteristic, in the training phase. Calculate the feature vector for the characteristics of the given character, and then calculate the resulted euclidean distance values over all characters generated in the database. The character with the nearest value will be the recognized character.

6.4.1 Definition of Euclidean distance:

Euclidean Distance Metric:

The formula for this distance between a point X (X1, X2, etc.) and a point Y (Y1, Y2, etc.) is:

$d = [\sum K\ (XK - Yk\ )\ 2]\ \frac{1}{2}$ ………………. (E.1)

Deriving the Euclidean distance between two data points involves computing the square root of the sum of the squares of the differences between corresponding values. The following fig. 6 illustrates the concept of euclidean distance:

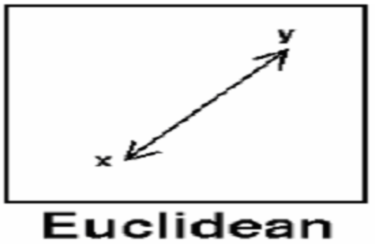


Fig.6: Euclidean distance calculation.

6.4.2 Recognition algorithm:

The nearest algorithm works like this: First the distance is calculated between all characters in the database and the test character. The distance measure that is used is the euclidian algorithm. After that, euclidean distance is calculated between test character and all database characters, and the character that gets the lowest value is said to be equal to the tested character. That is the characters that have the smallest distance is selected as recognize character.

Here we use Nearest Neighbor Algorithm for recognition. Take the vector you get from the unknown and compute its distance from all the patterns in your database by using Nearest Neighbor Algorithm and the smallest distance give the best match.

The Nearest Neighbor Algorithm is as

$$\boxed{D\,(a,\,b) = [\textstyle\sum_K (a_K – b_k)^2]^{1/2} \quad ;\ \text{for } k=1 \text{ to } d}$$

Where d= dimension.

6.4.3 Basic of nearest neighbor classification:

This algorithm is basic in design and implementation and yet is also widely used within machine learning and can provide good results. The idea behind the algorithm is that an item x is most likely to be of the same class as that of the item to which it is the most similar. Similarity in this instance is often defined as the euclidean distance metric between the two vector representations. The euclidean distances between two points x and y is given by equation (E.1).

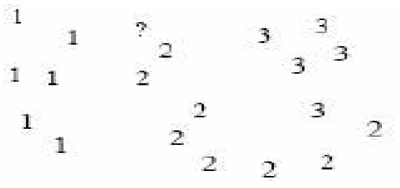


Fig.6: Nearest Neighbor Classification.

7. PERFORMANCE OF NEAREST NEIGHBOR ALGORITHM

We implemented different variations of the nearest neighbor algorithm and here we present some results. The first variant with larger databases uses a euclidean distance to calculate the nearest neighbor and takes the most common class as the answer. The other variants with smaller database also use the euclidian distance but the answer is dependent on how close these neighbors are. The test was performed with databases of different sizes, to see how this affects the performance. The test set consisted of samples not included in the learning set. For all these tests there is an increase in performance with larger databases. Test 1 resulted in a maximum success rate with the largest database. There was no significant difference between test 1 and the others.

It's a little bit strange that test 2 didn't perform better then test 1, but that is probably due to the simple distance measurement that considers irrelevant information such as line thickness and rotation etc. When only a couple of pixels differed between the unknown character and the reference, the results were fairly good, but larger differences often made the algorithm unable to correctly identify the unknown character. On the other hand, the low success rate is not indicative of the general algorithm, just the current implementation. There are many possible changes that could vastly improve the algorithm's recognition abilities. With a few of these changes implemented, the mistakes the algorithm would make would indeed be very similar to the types of mistakes humans would make. Thus general algorithm holds promise as a character recognizer that identifies characters in a manner similar to the way that humans identify characters.

8. RESULTS

Here are the tables which show the recognition results for database character and handwritten test character. The values given in the tables are the euclidean distance between trained character and test character. These values are obtained from the MATLAB program. During execution, the test characters are randomly taken. Suppose we consider calculation between handwritten character and database characters, having minimum ……………………………(E.3) euclidean distance with database characters as compare to rest characters. Therefore test handwritten character recognized as training or standard character

By observing these tables, test character having minimum euclidean distance value for the similar character in the training character column and for other characters it is maximum. Means either in rows or columns, block with minimum euclidean distance is the matching pair. These is the case satisfies for all similar character in all rows and columns. Hence we can say that test characters are mostly recognized.

9. FEATURES OF HCR SYSTEM

Handwriting recognition technology can provide benefits including: rapid generation of documents, productivity enhancements that free up more time for primary tasks at hand, reduction of transcription cost and delay, accuracy, mobility, ADA compliance, reduction of stress, and convenience.

9.1 Increased productivity:

Handwriting recognition technology allows users to get more done in less time, freeing them up to focus on the primary tasks at hand. This time is gained by dictating into computers, which produces text faster than by traditional keyboarding. Users can create, edit, and format documents, send e-mail, access and update records, and navigate their desktop and the Web.

9.2 Economy:

Eliminate the need for transcription services. Electronic and hardcopy transcriptions of dictated reports, letters, and other documents can be produced rapidly by a computer. In comparison, manual transcription requires additional costs and turnaround time, requires skilled transcriptions.

9.3 Speed:

Handwriting recognition generates text faster than keyboarding. That is text can be dictated successfully for automatic transcription at more words per minute than best typists with little or no physical effort. Using Handwriting enabled macro commands and application templates, the time required to perform usually lengthy tasks can be reduced to seconds.

9.4 Accuracy:

Handwritten notes can be minimized, along with the minimum possibility of misinterpretation. Transcribed text can be reviewed on screen as an accuracy check. Transcription accuracy for Handwriting recognition outputs can be very high. Users can train software in minutes to accurately recognize their handwriting. To facilitate accuracy, some products include extensive vocabularies that can be customized by users, and include specialized terms, phrases, and abbreviations. Typically, when software "misrecognizes" dictation, the result is syntactically obvious, and the misrecognition can be easily detected and corrected.

9.5 Mobility:

Computers with the digitizer and LCD screen are easy to handle and are more & more portable .The absence of keyboard & mouse helps to make computers more & more portable.

9.6 Reduced stress:

Handwriting recognition can provide significant physical and Psychological benefits. Documents can be produced in less time and with significantly less physical effort and stress than by keyboarding.

9.7 Convenience:

Handwriting recognition installed on a network allows users to dictate into any configured computer on the network, creating documents in real time virtually wherever they are in an enterprise. After transcription by the computer, an electronic record is available that can be immediately printed out, e-mailed, or linked to a Web page.

10. APPLICATIONS OF DHCR

This thesis describes an algorithm that attempts to work with a subset of the features in a character that a human would typically see for the identification of machine-printed Devnagri characters. Its recognition rate is as high as the recognition rates of the older, more developed character recognition algorithms, but it is expected that if it were expanded to work with a larger set of features this could be more advanced recognizer. If it were expanded to use more features, it would be made correspondingly slower; with the advent of faster microprocessors this fact is not viewed as a crippling problem. Another research area that is receiving a lot of attention now a day is the area of pattern recognition techniques for personal identification. Pattern recognition techniques can be used to protect PCs and networks from unauthorized access by authenticating user's base along with some other physical feature such as a fingerprint, retina, iris, hand, or face. Although voice and signature identification do not involve physical characteristics, they are usually included with pattern recognition techniques. Hence the area of application for handwritten character recognition is as given below,

- In ministry of documentary analysis.
- License checking in RTO office.
- Passport verification in airline services.
- Signature analysis in banking.
- Various field of handwritten word recognition.
- Handwriting Analysis Environment.

11. CONCLUSION

The above-discussed method has all the characteristics that are required for an off-line handwritten character system. That is simplicity and shortness of calculations. The method was found to be extremely reliable in preliminary dataset. The method can easily be applied to any application that requires handwritten character recognition, regardless of its computing power. This is due to the low computational requirement. Thus, the proposed algorithm can be implemented on any type of software platform. The method can also be applied to an on-line system if the coordinate data sent into the system can be sent in as a time ordered sequence of data.

Despite variations in character size, orientation, and position, the pattern recognizer system was still able to recognize many of the characters. While 2D image recognition is only part of the solution pattern recognition can bring to handwriting character recognition. Combined with euclidean distance analysis and temporal information, pattern recognition look to be a very promising solution. The results of this project are hardly new.

While pattern recognition is a promising solution there are some short-term problems. Conducting experiments on this project, it became clear that correctly training character for database; it can be a very time and processor intensive activity. This has resulted in some researchers advocating 'lazy recognition' that is offline recognition, not attempting to do character recognition in real time. Microsoft has also adopted a strategy with its upcoming Tablet PC to do character recognition silently in the background instead of in real time.

12. REFERENCES

[1] N. Sharma, U. Pal, F. Kimura, S. pal, "Recognition of Off Line Handwritten Devnagari Characters using Quadratic Classifier", ICCGIP 2006, LNCS 4338, 2006.

[2] R. Kapoor, D. Bagai and T.S. Kamal, "Representation and Extraction of Nodal Features of DevNagri Letters", Proceedings of the 3rd Indian Conference on Computer Vision, Graphics and Image Processing.

[3] Reena Bajaj, Lipika Dey, and S. Chaudhury, "Devnagari numeral recognition by combining decision of multiple connectionist classifiers",Sadhana, Vol.27, part. 1, , 2002

[4] Bellili, M. Gilloux, P. Gallinari, An MLP-SVM combination architecture for o²ine handwritten digit recognition: reduction of recognition errors by support vector machines rejection mechanisms, Int. J. Document Analysis and Recognition, 5(4): 2003.

[5] L. Xu, A. Krzyzak, C.Y. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, IEEE Trans. System Man Cybernet.,22, 1992.

[6] S. Arora, D. Bhattacharjee, M. Nasipuri, M.Kundu, D.K. Basu, "Application of Statistical Features in Handwritten Devnagari Character Recognition", International Journal of Recent Trends in Engineering[ISSN 1797-9617], IJRTE Nov 2009

[7] U. Kreßel, J. Schurmann: Pattern classification techniques based on function approximation. In: H. Bunke,P.S.P. Wang (eds.), Handbook of Character Recognition and Document Image Analysis, World Scientific, Singapore, 1997,

Author Biographies

Mr. Anil L. Telang had completed Graduation in Industrial Electronics in 2003 & pursuing the Post Graduate Degree (M.E.) in Digital Electronics from S.G.B. Amravati University, Amravati. Currently he is working as a Lecturer in Electronics & Telecommunication Department at Anuradha Engineering College, & Having teaching Experience of 7 Years. Chikhli. He is Presented 1 Paper in National & International Conference & 5 Papers in national level paper presentation.

Mr. Nagesh B. Mapari had completed Graduation in Electronics & telecommunication in 2005 & pursuing the Post Graduate Degree (M.E.) in Digital Electronics from S.G.B. Amravati University, Amravati. Currently he is working as a Lecturer in Electronics & Telecommunication Department at Anuradha Engineering College, & Having teaching Experience of 1 Years. Chikhli. He is Presented 1 Paper in International Conference & 2 Papers in national level paper presentation.

Mr. Ravi K. Rajbhure had completed the Bachelor of Degree from Information Technology Branch in the year 2006 from S.G.B. Amravati University and pursuing Post Graduation (M.Tech) in Information Technology from RGPV University at PCST, Indore. Currently he is working as a Lecturer in Information Technology Department at Anuradha Engineering College, Chikhli. & Having teaching Experience of 1 Years. His interest is in Embedded System and Real Time Application. He is Presented 3 Paper in International Conference & 4 Papers in National level Conference.