# A Benchmark Dataset for Devnagari Document Recognition Research

RAJIV KUMAR, AMRESH KUMAR, PERVEZ AHMED
Intelligent System Research Group
Sharda University
Greater Noida
INDIA
rajiv.kumar@sharda.ac.in http://www.sharda.ac.in

*Abstract:* - A benchmark dataset is required for the development of an efficient and a reliable recognition system. Unfortunately, no comprehensive benchmark dataset exists for handwritten Devnagari optical document recognition research, at least in the public domain. This paper is an effort in this direction. In here, we introduce a comprehensive dataset that we referred to as CPAR-2012 dataset, for such benchmark studies, also present some preliminary recognition results. The dataset includes 35,000 isolated handwritten numerals, 83,300 characters, 2,000 constrained and 2,000 unconstrained handwritten pangrams. It is organized in a relational data model that contains text images along with their writer's information and related handwriting attributes. We collected the handwriting samples from 2,000 subjects who were chosen from different age, ethnicity, and educational background, regional and linguistic groups. The samples reflect expected variations in Devnagari handwriting. The digit recognition results using recognition schemes that uses simple most features & four neural network classifiers & KNN, and classifier ensemble have also been reported for benchmarking.

*Key-Words:* - Benchmark Dataset, Character Extraction, Devnagari Handwritten Numeral Recognition, K-Nearest Neighborhood Classifier (KNN), Neural Networks Classifiers.

## 1 Introduction

A large number of languages, dialects and scripts are being used in India. Among them Devnagari—the script of the national language (Hindi) —is the most widely used script. In such a multilingual environment, a computer-mediated system that can transcribe Devnagari text into the text of the regional languages and vice-versa would be a desired communication option. In such a heterogeneous community, a system that can transcribe unconstrained handwritten text of one language into another with acceptable accuracy would be an asset because it can facilitate communications from anywhere and at any time by capturing the data at its source. Thereby, such data capturing ability can pave the way for the development of many real life applications, like an automatic reading machine for visually challenged people. Undoubtedly, development of such applications is a formidable task as it needs an accurate and efficient optical character recognition system as a front-end system and equally efficient and intelligible text to speech generator as a back-end system. In practice, development of these systems poses many challenges. However, despite challenges, the social and humanitarian obligations and most importantly automated data capturing requirements have motivated scientists and engineers to put in their best efforts in building an acceptable character recognition system for almost all major languages [1,2] including Hindi language [3,4] and the effort is still on.

Initially, the focus of the research in character recognition was mainly on the development of feature extraction, selection and classification methods. Their performance was measured on small experimental datasets [9,12,14]. Often the test dataset consisted of either artificial or digitized character image samples. Therefore, the recognition reliability and robustness of such methods could not be considered of practical importance. To overcome these limitations, need for obtaining the experimental dataset from real-life handwriting sample was realized. Currently, several real-life datasets are available for benchmark studies [1-9]. These real-life handwriting samples are not only helping in measuring the reliability and robustness of the recognition systems but also providing insights into the phenomena that control handwriting attributes like stroke formation styles. Such insights, in turn, are helping in devising better recognition techniques [3]. This work is an attempt to create a benchmark dataset for Devnagari digital document research. As compared to Latin languages [4-6], efforts to collect benchmark dataset for Hindi language have been very minimal.

Table 1: Dataset available for Devnagrai Scripts and other Indian Languages

| Year Ref. | Language | Dataset Size | Dataset Type | Number of Writers |
|---|---|---|---|---|
| 2007 [7] | Devnagari/ Bangla/ Telugu/ Oriya/ Kannada/ Tamil | 22, 546/ 14, 650/ 2, 220/ 5, 638/ 4, 820/ 2, 690 | Isolated numerals | N/A |
| 2007 [8] | Bangla | 8,348 / 23,392 | Strings of Online numerals / Offline-Isolated Numerals | N/A |
| 2009 [9] | Devnagari | 22,556 | Isolated Numerals | 1,049 |
| 2010 [10] | Tamil and Kannada | 100,000 | Words | 600 |
| 2011 [11] | Kannada | 4,298 / 26,115 | Text lines / Words | 204 |
| 2011 [12] | Hindi & Marathi | 26,720 | Words of Legal Amounts | N/A |
| 2012 [13] | Bangla only & Bangla mixed with English | 100 & 50 | Words | N/A |

However, few noticeable efforts to create a realistic benchmark dataset for Indian languages have appeared in literature [7-13] recently, these are described in Section 2. In Section 3, we describe data collection process, storage format and data attributes. Section 4 describes the experiments with isolated numeral dataset. In Section 5 we present an analysis of our recognition results of initial experiments that we conducted to assess the recognition performance of standard techniques on this dataset. We conclude the paper with Section 6, where we summarize our observations and highlight our efforts of building an integrated research environment to support the sharing of benchmark dataset & algorithms, and processes for continuous data collection and storage from maximum possible sources.

## 2 Related Work

The test dataset creation for character recognition research of Indian languages is a recent effort [7-13,15]. Table 1 shows that serious efforts to collect Devenagri datasets began in the preceding decade [3, 5, 6] but these datasets are small as compared to the datasets of other major languages. Moreover, most of the datasets for Indian languages contain numerals only [7,9]. Surprisingly, none of them have Devnagari characters, words and texts with the writers' information. The writer's information is useful for handwriting analysis application development. However, the work of Nethravathi et al. [10] on design of Tamil and Kannada word recognition indicates that they had collected and compiled a large handwritten document dataset but that too did not incorporate writer's information. Experience shows, real-life handwriting samples are needed to build a realistic system. The study reveals the scarcity of such realistic datasets for Devnagrai character recognition research. For this purpose, we have developed a large dataset named as CPAR-2012 (Centre for Pattern Analysis and Recognition-2012). It contains samples that would help in understanding the complex structure of Devnagari characters, discovering features, training and testing the system in real environment.

As compared to other datasets, the CPAR-2012 dataset provides more information. It is much larger than the largest dataset of Devnagari digits as reported in Bhattacharya and Chaudhary [5]. Moreover, CPAR-2012 has a variety of data samples including digits, characters, and text images. In addition to these, it contains writers' information and color images. The color images are in a format that reflects the worst image representation scenario that may be expected in real life applications. Hence, it would help in a realistic assessment of the robustness of the system.

## 3 Data Collection

We collected data from writers belonging to diverse population strata. They belonged to different age groups (from 6 to 77 years), genders, educational backgrounds (from 3rd grade to post graduate levels), professions (software engineers, professors, students, accountants, housewives and retired persons), regions (Indian states: Bihar, Uttar Pradesh, Haryana, Punjab, National Capital Region (NCR), Madhya Pradesh, Karnataka, Kerala, Rajasthan, and countries: Nigeria, China and Nepal). Two thousand writers participate in this

experiment. In order to collect their personal information and handwriting samples we designed two forms. The first form referred to as Form-1, is for isolated numerals, characters and writer information collection as shown in Fig. 1. It has a header block that contains three fields: Text Type, Writer Number and Form Number followed by a machine printed Hindi character block used as specimen to write these characters in the space provided in the following block. The last block is reserved for writer information where every writer wrote his/her name, age, gender, education level, profession, writing instrument/color, designation, psychological condition, region and nationality. The second form that we refer to as Form-2 (see Fig. 4), is designed for constrained and unconstrained handwritten word collection from the pangram. We developed this pangram to study the handwriting variations using 13 most commonly used vowels, 14 modifiers and 36 consonants. The top portion of the Form-2 contains the pangram text specimen. All subjects were asked to write the pangram text on the guided line given below the pangram for constrained handwriting sample collection and repeat the same in the blank space (without guidelines) provided for unconstrained handwriting sample collection. We digitized the duly filled forms using HP Cano LiDE 110 scanner at resolution 300 dpi in color mode.

### 3.1 Form 1 processing

The isolated characters (numerals and alphabet) and writer information were extracted from the digitized image of Form 1. The extraction process begins with skew correction operation, if required, on Form-1. Afterwards it locates the machine printed character block, handwritten character block followed by writer's information block in the skew free images. To extract the image of a character the process locates the grid that contains a handwritten character in the character block, and performs the following steps.

1. Convert Form-1 image into a binary image using Otsu Method [15].
2. Remove noises (impression of other forms, salt and pepper noise, corner folding, physically damaged paper, extraneous lines, stapler pins marks) that might have occurred during the digitization process.
3. Perform hole filling morphological operations to obtain the uniform connected component.
4. Perform the labeling operation on the connected components obtained in step-3 to find the bounding box (top-left point, width and height) for each labeled region.

After step 4, we filtered out all those labeled components whose areas were less than a specified threshold. Under these criteria, 1,700 out of 2,000 forms were accepted and processed. In each accepted from 154 bounding boxes were detected, cropped, stored and displayed for verification see Fig. 2. Through this process, we accepted constrained handwritten samples of 15,000 numerals and 83, 300 characters. The process rejected poor quality samples (1400 samples). These samples have also been stored in the database for further investigation.


Fig. 1 Form-1 as used in isolated characters extraction.

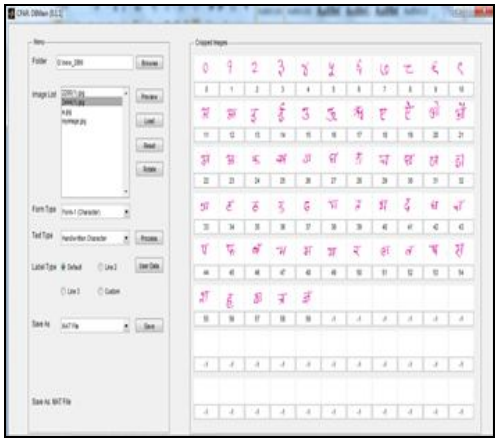


Fig. 2 Cropped character images from Form-1.

The writer information was extracted from the header block and writer information block of the Form-1. These two block images were cropped, combined and displayed for manual entry of the pertinent data. The collected data was saved in a relational database model for the purpose to access it using standard methods.

### 3.2 Form-2 Processing

Like before, we converted Form-2 images into binary images and removed the extraneous noises. The last line of each handwritten pangram (constrained and unconstrained) contains handwritten digits. We used three structuring elements that we referred to as SE1, SE2 and SE3 for image erosion operation. These elements are depicted in Fig. 3.

$$\underset{\text{SE1}}{\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{15\times 25}} \quad \underset{\text{SE2}}{[1 \quad \cdots \quad 1]_{1\times 100}} \quad \underset{\text{SE3}}{\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{100\times 25}}$$

Fig. 3 Structuring elements used in Form-2 processing

In order to extract individual digits from these lines following steps were performed.

1. Erode the binary image with structuring element SE1 so that isolated characters were merged within the word.
2. Erode the resultant image with a line structuring element SE2, that resulted image in a connected component.
3. Label the connected components obtained in step-2 to find region properties (top-left, width and height).
4. Select the last connected component and perform the following steps as a selection criteria:
   *(a)* If found acceptable once then save it, else go-to step (b)
   *(b)* Erode the same input image with another structuring element SE3 and check for acceptability condition as in (a) and go-to(c).
   *(c)* If acceptable then save the component and perform step 5, otherwise discard the component.
5. Invert the resulted component image from step- 4 and erode it to produce a list of white regions bounded by black regions.
6. Find the top left point, width and height of each white region, and that allowed to crop the individual digit image, store and display it for manual inspections and labeling.

Through this process, we collected 15,000 unconstrained and 5,000 constrained handwritten numerals. However, in case of constrained numerals, we lost some samples because writers wrote them over the upper, lower or both guide lines. The final dataset consists of: 83,300 isolated characters; 35,000 numerals; 2,000 constrained pangrams and 2,000 unconstrained pangrams; Writer's Information; 2,000 Form-1 images and 2,000 Form-2 images.

## 4 Experiments with Isolated Numerals of CPAR-2012 Dataset

We are in the process of conducting a set of benchmark studies on this dataset. In this section we present the recognition experiments that we have conducted, using the simplest feature, i.e., the direct pixel values. So, the recognition results of this experiment can form a basis to compare the results obtained by using more complex features. The feature vector of size 400x1 was formed by concatenating the columns of size normalized image of 20 rows and 20 columns.We tested the performance of the said feature using five different classifiers below:

1. Pattern Recognition network (PR): It is a specialized feed-forward neural network that uses tan- sigmoidal transfer function in the last layer for pattern recognition.
2. Feedforward Neural Network (FFN)
3. Function fitting neural network (FFT): It fits the input-output relationship through generalizing the non-linear relationship between example input and output.
4. Cascade Neural Network (CCN): It has a connection from input to all the layers.
5. KNN (k-nearest neighbor) classification.

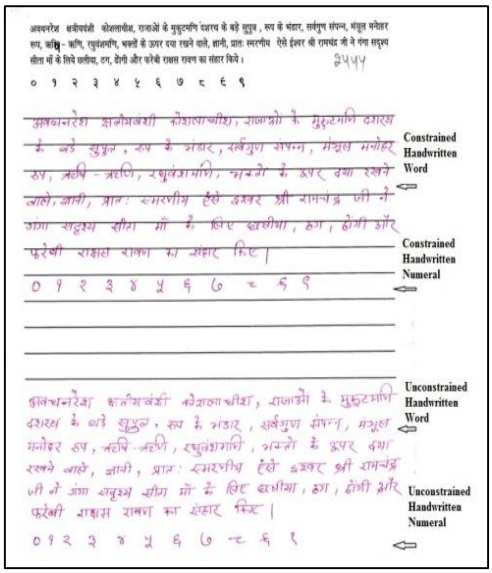



Fig. 4 Form-2 for constrained and unconstrained pangram extraction.

We experimented with the first four classifiers using scaled conjugate back-propagation (SCG) and resilient back-propagation (RP) learning algorithms.

Table 2: Results of different classifiers based on direct pixel method in four different experiments.

| Experiment Type | Training Type | Classifiers | | | | |
|---|---|---|---|---|---|---|
| | | PR | FNN | FFT | CCN | KNN |
| I | SCG | **88.6 (C2)** | **88.2 (C3)** | 86.5 | **87.1 (C5)** | **92.49 (C1)** |
| | RP | 86.8 | **87.2 (C4)** | 86.3 | 86.2 | |
| II | SCG | 92.5 | 90.2 | 90.5 | 89.1 | 100 |
| | RP | 88.9 | 89.2 | 89.3 | 89.2 | |
| III | SCG | 93.5 | 91.2 | 91.5 | 90.1 | 100 |
| | RP | 81.0 | 90.2 | 90.3 | 90.2 | |
| IV | SCG | 85.7 | 86.3 | 84.2 | 85.8 | 89.9 |
| | RP | 81.4 | 84.8 | 89.8 | 81.3 | |

In order to train and test all the above classifiers, we divided the dataset into two sets: A and B respectively. The set A has 11,000 and set B 24,000 digit samples taken from the CPAR-2012 dataset. Table 3 shows the distribution of digit samples of CPAR-2012. We conducted four experiments, that were referred to as experiment I, II, III and IV. In experiments I, the system was trained on dataset B and tested on A, likewise in experiments II , III and IV we trained and tested the system on the dataset B, trained and tested on the dataset A, and trained on the dataset A and tested on B respectively.

## 5 Experimental Results Analysis

This section discusses the validity of our benchmark dataset. We conducted recognition experiments using MATLAB-2009 on Intel Core 2 Duo 2.00GHz based system with 4 GB internal memory. Table 2 summarizes the experimental results. The results indicate that the recognition rate varies from 80% to 100%. The results further indicate that the SCG neural network based classifiers performed better in all the experiments. Among them, the PR network using SCG yielded the best result. It is noticeable that the KNN classifier has yielded the best recognition result but at the expense of a set of a large number of prototypes. One worthwhile observation is that the accuracy of recognition results is improved by increasing the training sample size (see columns in Table 2). It is widely claimed that use of a classifier ensemble should improve the recognition accuracy. To verify the claim, we conducted an experiment by combining the classifiers decisions using majority voting strategy.

Table 3: The CPAR-2012 Numeral dataset

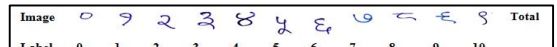
| Image |  | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Label | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Taining | 2280 | 2280 | 2280 | 2280 | 2280 | 2280 | 2280 | 2280 | 2280 | 2280 | 1200 | 24000 |
| Test | 1012 | 1012 | 1012 | 1012 | 1012 | 1012 | 1012 | 1012 | 1012 | 1012 | 880 | 11000 |
| Total | 3292 | 3292 | 3292 | 3292 | 3292 | 3292 | 3292 | 3292 | 3292 | 3292 | 2080 | 35000 |

Table 4. Confusion matrix after applying majority voting

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9(1) | 9(2) | Total Samples | %age Recognition Results |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 987 | 5 | 4 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 3 | 1010 | 97.72 |
| 1 | 0 | 980 | 5 | 0 | 3 | 3 | 2 | 6 | 0 | 3 | 7 | 1009 | 97.13 |
| 2 | 0 | 3 | 989 | 3 | 0 | 4 | 1 | 0 | 3 | 7 | 1 | 1011 | 97.82 |
| 3 | 0 | 5 | 33 | 940 | 3 | 4 | 1 | 6 | 8 | 2 | 5 | 1007 | 93.35 |
| 4 | 0 | 2 | 8 | 2 | 954 | 12 | 3 | 7 | 12 | 7 | 3 | 1010 | 94.46 |
| 5 | 0 | 12 | 26 | 8 | 30 | 914 | 8 | 3 | 0 | 9 | 0 | 1010 | 90.50 |
| 6 | 0 | 6 | 6 | 4 | 8 | 5 | 934 | 11 | 12 | 19 | 6 | 1011 | 92.38 |
| 7 | 22 | 6 | 0 | 0 | 6 | 4 | 10 | 957 | 0 | 6 | 0 | 1011 | 94.66 |
| 8 | 0 | 0 | 11 | 2 | 3 | 0 | 4 | 0 | 986 | 5 | 0 | 1011 | 97.53 |
| 9(1) | 0 | 0 | 16 | 6 | 2 | 2 | 14 | 4 | 2 | 959 | 6 | 1011 | 94.86 |
| 9(2) | 0 | 14 | 9 | 0 | 5 | 0 | 5 | 0 | 4 | 0 | 842 | 879 | 95.79 |
| | | | | | | | | | | | | 10980 | **95.11** |

For majority voting scheme, five best classifier values were selected from the experiment I (highlighted in bold characters in Table 2). The classified values of these classifiers formed the basis for majority voting. An unknown digit is recognized as the one which is recognized by the majority of classifiers. In case of a tie, a weighted voting mechanism was followed. The tie was resolved by aggregating the weights of each group of classifiers and applying the rule: Recognize the unknown digit as the one supported by the group having the greater weight, otherwise recognize as the one that is supported by the classifier that has the maximum weight among all the classifiers.If all the classifier doesn't agree on common consensus then the sample is rejected. The majority voting classifier yielded 95.11 % which is higher than 92.4% recognition rate yielded by the best performing classifier which is KNN classifier as shown in Table 4.

## 6 Conclusion

In this paper, we have presented a benchmark study on handwritten Devnagari digit recognition that we have collected from a large heterogeneous writers' groups. The dataset contains digits, characters and words for recognition and text for handwriting analysis. It is the largest dataset that has been collected in a real life writing environment for research in Devnagari optical document recognition research. The salient features of the dataset are: it has 35,000 digits; 83,300 characters; 2,000, constrained handwritten pangram images; 2,000 unconstrained handwritten pangram images; writer's information; and original images of data collection forms. The CPAR-2012 dataset is available in the public domain. The dataset can be accessed through the Integrated Research Environment for Devnagari optical Document Recognition hosted at www.cpar.co.in. In these experiments, the KNN classifier yielded 100% recognition rate on the training set and 92.49% on the test set. The best neural network classifier (Pattern Recognition) yielded 88.60% on test set. In an attempt to study the effect of classifier ensemble we used majority voting classifier combination strategy that improved the recognition rate to 95.11% on the test set. More experiments are underway, on this dataset, with the best performing recognition schemes that are reported in Devnagari digital document processing literature. The findings of these experiments shall be made available once results get authenticated.

*References*:


[1] C. Y. Suen, M. Berthod, and S. Mori, Automatic recognition of handprinted characters– The state of the art, *Proceedings of the IEEE* Vol. 68, No. 4, 1980, pp. 469–487.

[2] Al–Ohali, Yousef, M. Cheriet and C. Y. Suen, Databases for recognition of handwritten Arabic cheques, *Pattern Recognition*, Vol. 36, No. 1, 2003, pp. 111-121.

[3] C. Y. Suen, C. Nadal R. Legault, T. A. Mai, and L. Lam, Computer recognition of unconstrained handwritten numerals, *Proceedings of the IEEE*, Vol.80, No. 7, 1992, pp. 1162–1180.

[4] J. J. Hull, A database for handwritten text recognition research, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.80, No. 7, 1992, 1162–1180.

[5] W. Andrew Senior, and A. J. Robinson, An off-line cursive handwriting recognition system, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, 1998, pp. 309–321.

[6] LeCun, Yann, and C. Cortes, The MNIST database of handwritten digits, URL: http://yann. lecun. com/exdb/mnist, 1998

[7] U. Pal, T. Wakabayashi and F. Kimura, Comparative study of Devnagari handwritten character recognition using different feature and classifiers, *In 10th International Conference on Document Analysis and Recognition*. ICDAR'09, 2009, pp. 1111–1115.

[8] B. B. Chaudhuri, A complete handwritten numeral database of Bangla-a major Indic script, *In Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

[9] U. Bhattacharya and B. B.Chaudhuri , Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 3, 2009, pp. 444–457.

[10] B. Nethravathi,, C. P.Archana, , K. Shashikiran, A. G.Ramakrishnan, and V.Kumar , Creation of a huge annotated database for Tamil and Kannada OHR, *In 2010 IEEE International Conference on Frontiers in Handwriting Recognition*, (ICFHR), 2010, pp. 415–420.

[11] Alaei, Alireza, , P.Nagabhushan , and, U.Pal , A benchmark Kannada handwritten document dataset and its segmentation, *In 2011 IEEE International Conference on Document Analysis and Recognition*, (ICDAR) 2011, pp. 141–145.

[12] R.Jayadevan, , S. R.Kolhe, , P. M.Patil, , and U.Pal, Database development and recognition of handwritten devanagari legal amount words, *In 2011 IEEE International Conference on Document Analysis and Recognition*, (ICDAR) , pp. 304-308.

[13] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, CMATERdb1: a database of unconstrained handwritten Bangla and Bangla English mixed script document image, *International journal on document analysis and recognition*, Vol. 15, No. 1, pp 71-83 2011, pp. 1–13.

[14] R. Jayadevan, S. R. Kolhe, P. M. Patil and U. Pal, Offline recognition of Devanagari script: A survey, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 41, No. 6, 2011, pp. 782–796.

[15] Ostu, Nobuyuki, A threshold selection method from gray-level histogram. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 9, No. 1, 1979, pp. 62-66.