# Brief review of research on Devanagari script

**Holambe A.N.[1]*, Thool R.C.[2], Shinde U.B.[3] and Holambe S.N.[4]**
*[1,4] Department of Information Technology, College of Engineering, Osmanabad-413501, MS, India, anholambe@yahoo.com.
[2]Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering & Technology, Vishnupuri, Nanded-431606, MS, India, rcthool@yahoo.com
[3]Savitribai Phule Women's Engineering College, Aurangabad, MS, India, shindeulhas1@yahoo.com

**Abstract-** In this paper we have done a brief survey of Devanagari script, research and different classifiers approach used for Character Recognition .We have collected and created database of handwritten characters. We have preprocessed it, extracted feature using local intensity distribution of gradient and used in our experiment. We have used KNN Classifier. Experiment result illustrates the accuracy, rejection, error percentage of classification.
**Keywords-** Devanagari script, KNN Classifier, accuracy, rejection, error percentage of classification

## Introduction

Character recognition, is known as OCR (Optical Character Recognition) is an area within the pattern recognition. Optical Character Recognition deals with automatic recognition of different characters in a document image leading to clear and unambiguous recognition, analysis and understanding of the document content. The task of recognition can be broadly separated into two categories: machine printed data and the handwritten data. Machine printed characters are consisting of font used by user. They are unique and uniform. While handwritten characters are non-uniform; there size, shape depends on writer and the pen used by the writer. Handwriting of same writer may vary depending on the situation in which he is writing.

## Devanagari Script

The name Devanagari comes from the Sanskrit words Deva (god), and Nagari (city); together they mean, literally, the script of the "City of the Gods", where this city is the body of the individual. Chronological development of the script from the early Brahmi to the modern day [26]. Devanagari is indicated in table 1 and “fig.” 1,(Courtesy www.iitm.ac.in [26]). There are about a thousand conjunct consonants, most of which combine two or three consonants. There are also some with four-consonant conjuncts and at least one well-known conjunct with five consonants. Devanagari has no case distinction, i.e. no majuscule and minuscule letters.


Fig. 1-

Table 1- *Stages In The Evolution Of The Script*.

| | |
|---|---|
| 300 BCE | Mauryan : Early Brahmi form the Asokan edicts. Some scholars believe that Brahmi itself evolved from "karoshti" a script written right to left. |
| 200 CE | Kushan/ Satavahana Dynasties. |
| 400 CE | Gupta Dynasty |
| 600 CE | Yasodharman |
| 800 CE | Origins of the present day Nagari Script. Vardhana dynasty in the North and Pallava period in the South. |
| 900 CE | The period of the Chalukyas and Rashtrakutas |
| 1100 CE | Continuation of the Chalukya Rule |
| 1300 CE | Yadavas in the north and Kakatiyas in the south. |
| 1500 CE | The Vijayanagar empire. |

Most of the Indian scripts are originated from Brahmi script through various transformations. Among Indian scripts, Devanagari is the most popular script in India and the most popular Indian language Hindi is written in Devanagari script. Nepali, Sanskrit and Marathi are also written in Devanagari script. Moreover, Hindi is the national language of India and the third most popular language in the world. Devanagari script consists of a set of vowels and consonants along with various modifier symbols.Writing style in the script is horizontal, left to right and the characters do not have any uppercase/lowercase distinction. There are about fifty basic characters in scripts having nearly one to one correspondence. Within a word, the vowel characters often take modified shapes called modifiers. Consonant modifiers are also possible. Moreover, between two to four consonants can combine to form near about 250

compound characters, which partly retain the shape of the constituent consonants. In addition, most of the basic and compound characters can be attached with modifiers to generate new shapes. Apart from these, the documents printed in these scripts show large variations in font faces, type styles, and in character sizes. For a large number of characters it may be noted that there exists a horizontal line at the upper part. This line called shirorekha in Devanagari and is referred here as headline. The neighboring characters of a word very often touch through the headline to form a connected component. In script, a text word may be partitioned into three zones. The upper zone denotes the portion above the headline (ascenders); the middle zone covers the main portion of the basic and compound characters and the lower zone, where some vowel and consonant modifiers (descenders) can reside. Devanagari vowels are not scattered in the 'Varnamala' but are arranged at the beginning of the alphabet.

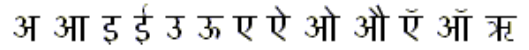


Fig. 2-

Devanagari Vyanjans

These are very logically arranged in following groups.

A.Sparsh

B.Antashth

C. Ushm

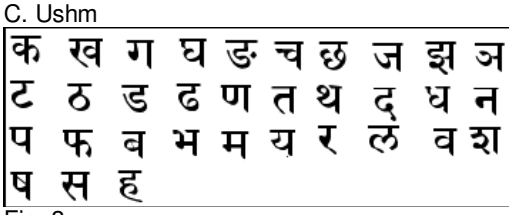


Fig. 3-

In the following figure you can see swars (vowels and diphthongs) and their respective matras. The dotted circles represent a placeholders for consonants (or base-letters), so we can understand the relative position of these matras.

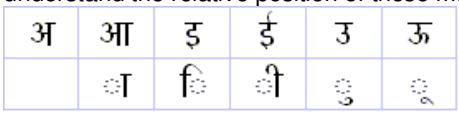


Fig. 4-

shuddh _vyaNjan (Half Form)

shuddh vyaNjan means pure consonants. These are those vyaNjan which are pronounced without the inherent vowel. So while transliterating these vyaNjan we won't write trailing . While typing Nagari we get these pure consonants by typing halant after the regular form. The aakaar (vertical line) is generally removed when we write these shuddh vyaNjan.

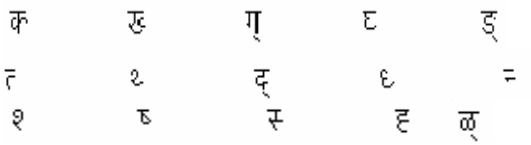


Fig. 5-

Nuktaa _vyaNjan

Nuktaa is a diacritic mark. Following two

vyaNjan are very common in Hindi.

chandra-bindu Chandra-bindu is a nasalization mark.

With ि, ी, े, ै, ो and ौ instead of ँ (chandra-bindu) people generally write ं (Anusvaar); i.e. instead of किं, कीं, कें, कैं, कों and कौं people generally write कि, की, कें, कैं, कों and कौं.

joDda_AkShar

joDda means join or joint. So combined letters (conjuncts) are called joDda_AkShar.

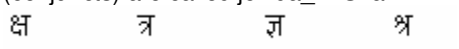


Fig. 6-

Devnagari Numerals

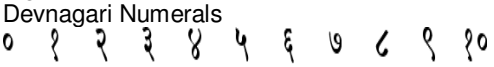


Fig. 7-

**Devnagari Script and Research**

OCR work on printed Devanagari script started in early 1970s. Among the earlier pieces of work, some of the efforts on Devanagari character recognition are due to Sinha [1,7,8] and Mahabala [1]. Sethi and Chatterjee [5] also have done some earlier studies on Devanagari script and presented a Devanagari hand-printed numeral recognition system based on binary decision tree classifier. They [6] also used a similar technique for constrained hand-printed Devanagari character recognition. They did not show results of scanning on real document pages. The first complete OCR system development of printed Devanagari is perhaps due to Palit and Chaudhuri [4] as well as Pal and Chaudhuri [3]. For the purpose some standard techniques have been used and some new ones have been proposed by them. The method proposed by Pal and Chaudhuri gives about 96% accuracy. A survey for hand-written recognition of character is proposed [2]. A few of these work deals with handwritten characters of Devanagari. Because of the complexities involved with Devanagari script, already existing methods can not be applied directly with this script report on handwritten Devanagari characters was published in 1977 [9] and not much research work is done after that. Some research work are available towards Devanagari numeral recognition [10-12] but to the best of our knowledge there are only two reports on Devanagari off-line handwritten character

recognition [13,14] after the year 1977. One work is due to Kumar & Singh [13] and they proposed Zernike moments based approach for Devanagari character recognition. The other work on Devanagari character recognition is proposed [14] and 64 dimensional chain code features have been used in that work, but still no any standard OCR is available for the same. An excellent survey of the area is given in [15].For recognition of handwritten Devanagari numerals, Ramakrishnan et al. [16] used independent component analysis technique for feature extraction from numeral images. Bajaj et al [11] considered a strategy combining decisions of multiple classifiers. In all these three studies, very small sets of samples were considered. In an attempt to develop a bilingual handwritten numeral recognition system, Lehal and Bhatt [17] used a set of global and local features derived from the right and left projection profiles of the numeral images for recognition of handwritten numerals of Devanagari and Roman scripts.

**The OCR Data Set Creation**

We have collected handwritten characters from different peoples of different age group (i.e. 03 to 75), i.e. of 7000 people of different age groups, the datasheet used for collection is shown in figure 7. We have visited schools, High school, colleges, Government offices, Adult education schools for the Data collection.

While preprocessing we have consider the distortion in image because of users pen and writing quality. We have performed Preprocessing operations for rectification of distorted images, improving the quality of images for ensuring better quality edges in the subsequent edge determination step.

In order to remove noise and diminish spurious points, Which are introduced by uneven writing surface, we have used filtering operation. we have performed smoothing, sharpening, thresholding and contrast adjustment by using filtering operation[22][23].we have applied skew Normalization, slant Normalization, size normalization ,curve smoothing, to remove all types of variations during the writing and obtain standardized data[24].Then we have performed thinning operation to get better features[25].

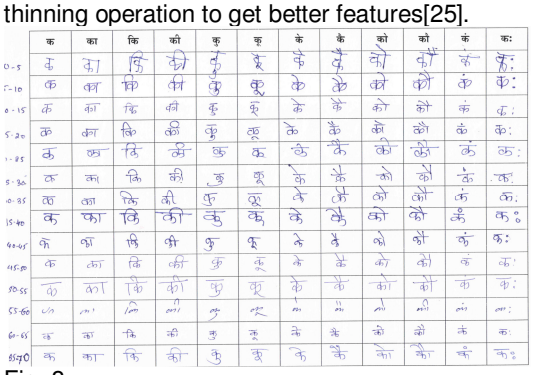


Fig. 8-

**Feature Extraction**

We have extracted feature on our dataset, the feature set consisted of local intensity distribution of gradients computed on 5 X5 grid using 16 quantized gradient orientations [18,19]. The original binary image was converted to a gray image using Gaussian filter, here feature vector size is 400 .The number of blocks are initially 9 X 9 and down sampled to 5 X 5.Here we are using 16 direction level .In order to obtain 16 directions Gaussian filter and a Robert filter are applied to the character image to obtain a gradient image. The arc tangent of the gradient is quantized into 16 directions and the strength of the gradient is accumulated in each direction in each block.

**Euclidian Distance-Based K-NN Classification**

In KNN classification, training patterns are plotted in d-dimensional space, where d is the number of features present. These patterns are plotted according to their observed feature values and are labeled according to their known class. An unlabelled test pattern is plotted within the same space and is classified according to the most frequently occurring class among its K-most similar training patterns; its nearest neighbors. The most common similarity measure for KNN classification is the Euclidian distance metric, defined between feature vectors as:

$$euc(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{f} (x_i - y_i)^2} \quad (1)$$

Where f represents the number of features. Smaller distance values represent greater similarity [20, 21].

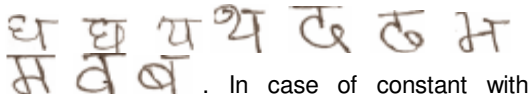
**Results**

We have used 20,000 Hand written sample data files. We have used 1200 for testing. We have organized data in class, separate class is decided for each character (i.e. vowels ('svar'), consonants ('vyanjan') without modifiers, consonants ('vyanjan') with modifiers(%)).Then we have taken result. Our results are given in the table 1.we have computed accuracy of each individual Devanagari Character the table shows the average accuracy.

*Table 2- Average Result Of Our Dataset.*

| | Accuracy | Rejection | Error |
|---|---|---|---|
| vowels ('svar') | 98% | 3% | 2% |
| consonants ('vyanjan') | 97.50% | 6% | 5% |
| without modifiers | | | |
| consonants ('vyanjan') | 94% | 7% | 9% |
| with modifiers(%) | | | |

In case of we get high accuracy and rejection and error is also less because they all are unique. But in case of constants we high rejection and error

because of these characters  . In case of constant with modifiers the error and rejection is high because of the writing style of writer and combine characters.

**Future Scope**

We have to develop a standard handwritten database for identifying the script.We have to work for combined letters (conjuncts) are called joDda_AkShar, shuddh _vyaNjan (Half Form) Half-form: those which joins to the next consonant. Two or more classifiers should be combined by different approach of classifier combining should be used, so that the identifying accuracy should be improved. A generalized Character Classifiers should be designed by combining different Classifiers which works on printed OCR as well on Handwritten Characters.

**References**


[1] Sinha R.M.K., Mahabala H. (1979) *IEEE Trans. System, Man Cybern.*, 9, 435-441.

[2] Plamondon R. Srihari S.N. (2000) *IEEE Transactions On Pattern Analysis And Machine Intelligence*. 22(1), 63.

[3] Pal U., Chaudhuri B.B. (1997) *Vivek*, 10, 12-24.

[4] Palit S., Chaudhuri B.B., Das P.P., Chatterjee B.N. (1995) *(Eda.) Pattern Recognition, Image Processing and Computer Vision, Narosa Publishing House: New Delhi, India*, 163-168.

[5] Sethi K., Chatterjee B. (1976) *J. Inst.Electron. Telecom. Eng*. 22, 532-535.

[6] Sethi K., Chatterjee B. (1977) *Pattern Recognition* 9, 69-76.

[7] Sinha R.M..K. (1973) *Ph.D. Thesis , Electrical Engineering Department, Indian Institute of Technology, India*.

[8] Bansal V., Sinha R.M.K. (1999) *Proceedings of the Fifth International Conference on Document Analysis and Recognition* , 653-656.

[9] Arora S., Bhattacharya D., Nasipuri M., Malik L. (2006) *IEEE –International Conference on Signal And Image Processing, Hubli, Karnataka*.

[10] Hanmandlu M. and Ramana Murthy O.V. (2005) *In Proc. Intl. Conf. on Cognition and Recognition*, 490-496.

[11] Bajaj R., Dey L., and Chaudhury S. (2002) *Sadhana,* 27, 59-72.

[12] Bhattacharya U., Parui S. K., Shaw B., Bhattacharya K. (2006) *In Proc. 10th IWFHR*, 613-618.

[13] Kumar S. and Singh C. (2005) *In Proc. Intl. Conf. on Cognition and Recognition*, 514-520.

[14] Sharma N., Pal U., Kimura F. and Pa S. (2006) *In Proc. Indian Conference on Computer Vision Graphics and Image Processing*, 805-816.

[15] Pal U. and Chaudhuri B.B. (2004) *Pattern ecognition*, 37, 1887-1899.

[16] Ramakrishnan K.R., Srinivasan S.H. and Bhagavathy S. (1999) *Proc. of the 5th ICDAR*, 414-417.

[17] Lehal G.S. and Nivedan Bhatt (2001) *Advances in Multimodal Interfaces–ICMI 2001*, Tan T., Shi Y. and Gao W. (2000) *(Editors), LNCS*, 1948, 442-449.

[18] Wakabayashi T., Tsuruoka S., Kimura F. and Miyake Y. (1995) *System and Computers in Japan*, 26 (8), 35-44.

[19] Kimura F., Miyake Y., Shridhar M. (1994) *Proc. Of 4Th IWFHR*.

[20] Cover T.M. and Hart P. E. (1967) *IEEE Trans. Inform. Theory*, IT-13, 21-27.

[21] Dasarathy B. V. (1991) *IEEE Computer Society Press, New York*.

[22] Lee J. S. (1983) *Digital Computer Vision, Graphics and Image Processing*, 24, 255-269.

[23] Haralick R. M. and Shapiro L. G. (1992) *Computer and Robot Vision,1, Addison _Wesley Publishing*.

[24] Guerfaii W. and Plamondon R. (1993) *Pattern Recognition*, 26 (3), 418-431.

[25] Lam L., Lee S. W. and Suen C. Y. (1992) *IEEE Trans. Pattern recognition and Machine Intelligence*, 14, 869-885.

[26] http://www.iitm.ac.in