A Survey on Devanagari Character Recognition for Indian Postal System Automation

Ankita S. Wanchoo  
Assistant Professor, Department of Electronics & Telecommunications, Symbiosis Institute of Technology, Affiliated to Symbiosis International University, Pune -412115, Maharashtra, India.

Preeti Yadav  
Assistant Professor, Department of Applied Science, Symbiosis Institute of Technology, Affiliated to Symbiosis International University, Pune- 412115, Maharashtra, India.

Alwin Anuse  
Department of Electronics & Telecommunications Maharastra Institute of Technology, Affiliated to University of Pune, Pune- 411038, Maharashtra, India.

Abstract
The commercial and industrial applications, namely, business form reading, bank cheque reading, and full postal address reading, have constituted the majority of the market for handwriting recognition technology. Devanagari is the most popular script in India and is used to write texts in Hindi, Marathi and Nepali languages. Significant amount of work has been done on Devanagari isolated characters. Very few attempts have been made for Indian postal system automation. This paper presents the challenges involved in Indian postal system automation with a case study. This paper also throws light on the existing research literature support available for doing the postal automation.

Keywords: Devanagri script, Character recognition

Introduction
Optical character recognition (OCR) based systems may be implemented in variety of applications such as reading handwritten digits from bank cheques, text extraction and recognition of hand filled structured form documents, automatic sorting of postal documents etc.[1]. In India, a multilingual multiscr ipt country, 18 official languages are written in 12 different scripts. Fewer than three language formulas, the documents for a region are written in English, Hindi (Devanagari) and the regional official language [2]. Recognition of numeric postal codes in a multi-script environment is a classical problem in any postal automation system [2].

In this paper, we present the challenges involved in Indian postal system with a case study of Pune city (Maharashtra state, India) area. The paper is organized as follows: Section 2 is dedicated to the characteristics of Devanagari script. Section 3 throws light on various considerations to be made in postal system automation with the challenges. Section 4 gives a brief survey of research work done on Devanagari character recognition. Section 5 gives the conclusions.

Characteristics of Devanagari Script
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 Hindi is the third most popular language in the world. Thus, the work on Devanagari script is very useful for the country. The alphabet of the modern Devanagari script consists of 14 vowels and 37 consonants. These characters may be called basic characters. The basic characters of Devanagari script are shown in Fig 1. Writing style in Devanagari script is from left to right. The concept of upper/lower case is absent in Devanagari script. In Devanagari script, a vowel following a consonant takes a modified shape. Depending on the vowel, its modified shape is placed at the left, right (or both) or bottom of the consonant. These modified shapes are called modified characters. A consonant or vowel following a consonant sometimes takes a compound orthographic shape, which we call as compound character. Compound characters can be combinations of two consonants as well as a consonant and a vowel. Compounding of three or four characters also exists in these two scripts. There are about 280 compound characters in Devanagari. A Devanagari text line can be partitioned into three zones. The upper-zone denotes the portion above the head-line, the middle zone covers the portion between head-line and base-line, the lower-zone is the portion below base-line [24].

Most of the Indian scripts are distinguished by the presence of matras (or, character modifiers) in addition to main characters as against the English script that has no matras. Therefore, algorithms developed for them are not directly applicable to Indian scripts. Many OCRs for Indian scripts have been reported [35-37]. However, none of these has attempted the handwritten Hindi text consisting of composite characters that involve both the main characters and matras [25].
Considerations and Challenges For Doing Indian Postal System Automation

1) First step is to acquire the image of postal document. Camera resolution is always going to play a major role in the recognition rate of the system.

2) Any Indian postal document might contain information in any language (18 official languages) written in 12 different scripts. The postal document of Pune region may have relevant information on it written in English, Hindi (Devanagari) and the regional official language which is Marathi (Devanagari). However, in other parts of the country the regional information may be in Tamil, Telugu, Kannada language etc. Challenge here lies in detecting the script for initial sorting of documents.

3) Postal document may be of different size, color etc.

4) Popular postal stationeries are printed by the Department of post having specific sizes, formats and structures. Postal documents having pre-determined address block structures are referred as structured postal documents [1]. But most of the business transactions takes place with the help of non-structured postal documents wherein there are no pre-determined address block structures. Challenge here lies in finding the script for initial sorting of documents. Fig 2 shows few out of many types of postal documents. It is easy to get pin code information from structured documents compare to non-structured documents.

5) The first step is to find the script of the document. The postal address may be in Devanagari, English or in both the languages for Maharashtra State (India). In a multiscr ipt, multilingual environment, it is essential to know the script used in writing document before an appropriate character recognition and document analysis algorithm can be chosen. Various Script identification techniques are available which are discussed in [27]. Gabor filter based texture feature with Fuzzy classifier is used for classification of Devanagari, Bengali, Telugu and Latin [28].

6) Next step is to find the addressee and the addresser before detecting the pincode as information of the addressee has to be processed for sorting the postal documents.

7) After finding the script and the addressee the next step is to find the pin code information on the document. For a structured document, it is very easy to find out the pin code information. For a non-structured postal envelope the problem of finding pin code is difficult because of the following things:

(a) Pin code may be written anywhere on the envelope (generally at the end).
(b) Pin code may be English, Devanagari or in both the languages. This demands bilingual processing.
(c) Table 1 shows various pin codes for Pune City (Maharashtra state, India). Bilingual processing of Devanagari and English numerals gives rise to some confusing situation as shown in Figure 3.
(d) Pin code written may not be complete i.e. it may have only last two digits or one digit
(e) Pin code written may be wrong.
(f) Pin code may be absent.
(g) Since the pin code written may be wrong, one cannot rely on pin code information for sorting. For correct sorting the address field has to found. From the address field the area can be found out which can be associate to pin code.
(h) Problems in using the area information for finding the pin code.
(i) Postal address information has to be segmented for separating area.
(j) Compound characters are difficult to recognize. A very small variation between two character leads to recognition complexity and degree of recognition accuracy. In Devanagari, there are many similar shaped characters. Examples of some groups of similar shaped characters are shown in Fig.4.
(k) Even if the words are segmented, the chances of them written wrongly cannot be ignored.
(l) Every area has sub areas, example: Yerawada in Pune city has Phulenagar, Vishrantwadi, Pratik Nagar, Mantle corner and Dhanori etc. The tendency in people is to write the sub area on the envelope.
(m) Two sub areas can have the same name e.g. Anandnagar in Kothrud (Pune) and Anandnagar in Singhad (Pune).
(n) Special characters like semicolon, colon, punctuation marks etc. are too considered.
Figure 2: Types of postal documents (a) Structured document, (b) Non structured document of white color, (c) Non structured document of brown color, (d) Envelope with address covered by plastic, (e) Structured document with address printed, (f) Inland Letter card with address printed, (g) Custom made card for Marriage invitation, (h) Envelopes for bank transactions.
Survey of Research on Devanagari Character Recognition Script

There are four major stages in the Character Recognition problem:
1) Preprocessing
2) Segmentation
3) Feature extraction
4) Training and Recognition

Preprocessing

The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Preprocessing aims to produce data that are easy for the character recognition system to operate accurately. The main objectives of preprocessing are 1) noise reduction; 2) normalization of the data; 3) compression in the amount of information to be retained [4]. The noise, introduced by the optical scanning device or the writing instrument causes disconnected line segments, bumps and gaps in lines, filled loops, etc. The distortion, including local variations, rounding of corners, dilation, and erosion, is also a problem. Prior to the Character Recognition, it is necessary to eliminate these imperfections.

Many noise reduction techniques are available which can be categorized in three major groups [6], [7]. The techniques are given in Table 2.

![Figure 3: Examples of some similar shaped numerals of Deva-nagari and English.](image3.png)

![Figure 4: Examples of Some Similar Shaped Devanagari Characters](image4.png)

**Table 1:** Pin codes of Pune city

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Category</th>
<th>Operation</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Filtering</td>
<td>Removes noise and diminish spurious points, usually introduced by uneven writing surface and/or poor sampling rate of the data acquisition device [4].</td>
<td>Filters can be designed for smoothing [8], sharpening [9], thresholding [11], removing slightly textured or colored background [10], and contrast adjustment purposes [12].</td>
</tr>
<tr>
<td>2</td>
<td>Morphological</td>
<td>Filter the document image replacing by the logical operations.</td>
<td>Morphological operations can be successfully used to remove the noise on the document images due to low quality of paper and ink, as well as erratic hand movement.</td>
</tr>
<tr>
<td>3</td>
<td>Noise Modeling</td>
<td>Noise could be removed by some calibration techniques if a model for it were available [4].</td>
<td>There is very little work on modeling the noise introduced by optical distortion, such as speckle, skew, and blur [13], [14].</td>
</tr>
</tbody>
</table>
Normalization
Normalization methods aim to remove the variations of the writing and obtain standardized data. There are four basic methods for normalization [15], [16].

a) Skew Normalization and Baseline Extraction: Due to inaccuracies in the scanning process and writing style, the writing may be slightly tilted or curved within the image. After skew detection, the character or word is translated to the origin, rotated, or stretched until the baseline is horizontal and retranslated back into the display screen space.

b) Slant Normalization: One of the measurable factors of different handwriting styles is the slant angle between longest stroke in a word and the vertical direction. Slant normalization is used to normalize all characters to a standard form [4]. However some of the recognition systems do not use slant correction and compensate it during training stage [17], [18].

c) Size Normalization: This is used to adjust the character size to a certain standard. The character is divided into number of zones and each of these zones is separately scaled [30]. Size normalization can also be performed as a part of the training stage, and the size parameters are estimated separately for each particular training data [19].

d) Contour Smoothing: It eliminates the errors due to the erratic hand motion during the writing.

Compression:
Two popular compression techniques are thresholding and thinning. In order to reduce storage requirements and to increase processing speed, gray-scale or color images are pixel wise or nonpixel wise thinning [31].

Segmentation
The preprocessing stage yields a “clean” document in the sense that a sufficient amount of shape information, high compression, and low noise on a normalized image is obtained. The next stage is segmenting the document into its subcomponents. Segmentation is an important stage because the extent one can reach in separation of words, lines, or characters directly affects the recognition rate of the script [4]. Complete review of these methods can be found in [20-23]. These text line segmentation methods usually make two assumptions:

1) The gap between two neighboring lines is significant, and
2) the lines are reasonably straight. However, these assumptions are not always valid for handwritten character recognition. Xiao JunDu, Wumo Pan, Tien D.Biu proposed a new text line segmentation method based on Mumford-Shah model. The specialty of this method is that it is script independent. Also an attempt is made to remove overlaps between neighboring text lines and connect broken ones using morphing [32].

Feature Extraction and Recognition Engine
Table 3 gives the summary of important feature extraction and classification methods used for Devanagari character recognition.

Table 3: Feature Extraction Algorithms for Devanagari Character Recognition

<table>
<thead>
<tr>
<th>NAME OF THE ALGORITHM</th>
<th>DESCRIPTION</th>
<th>CLASSIFICATION RATE</th>
<th>NEURAL NETWORK</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADIENT DIRECTION HISTOGRAM</td>
<td>Gradient of the character image is calculated using different operators: Sobel, Kirsh, Roberts and LoG. Kirsh masks directly calculate the gradient magnitude of four orientations whereas Roberts operator calculates two gradient components in orthogonal directions.</td>
<td>MLP: 99.69%</td>
<td>MLP (Multilayer Perceptron)</td>
</tr>
<tr>
<td>BOX APPROACH</td>
<td>Each character is divided into 24 boxes. By dividing the sum of distances of all black pixels present in box with the total no. of pixels, a normalized vector distance is obtained.</td>
<td>95%</td>
<td>Fuzzy neural network</td>
</tr>
<tr>
<td>TEMPLATE MATCHING [29]</td>
<td>Character image is used as a feature vector. In this recognition state, the similarity between the character and each template is calculated. Template having the lowest dissimilarity is measure and is given a different class label.</td>
<td>NOT MENTIONED</td>
<td>MLP</td>
</tr>
<tr>
<td>UNITARY IMAGE TRANSFORM [29]</td>
<td>Unitary transform is applied to the image, obtaining reduction of features while retaining most of the information about the character shape.</td>
<td>NOT MENTIONED</td>
<td>MLP</td>
</tr>
<tr>
<td>ZONING [29]</td>
<td>An n x m grid is superimposed on the character image and for each of the n x m zones the average is computed giving a feature vector n x m.</td>
<td>NOT MENTIONED</td>
<td>MLP</td>
</tr>
<tr>
<td>GEOMETRIC MOMENT INVARIANTS [29]</td>
<td>In this method Hu’s absolute orthogonal moment invariants have been used which is invariant to translation, rotation, scale. These invariants are of order 2.</td>
<td>NOT MENTIONED</td>
<td>MLP</td>
</tr>
<tr>
<td>SPLINE CURVE APPROXIMATION [29]</td>
<td>In this high curvature approximation points called breakpoints on the outer character contour and approximates the curve between the two break points with the spline function. The both the break point and the spline curve parameters are used as features.</td>
<td>NOT MENTIONED</td>
<td>MLP</td>
</tr>
<tr>
<td>POSITIONAL FEATURES [43]</td>
<td>It determines the relative position of the global or geometrical reference points.</td>
<td>93.7%</td>
<td>NOT MENTIONED</td>
</tr>
<tr>
<td>Feature Description</td>
<td>Feature Values</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature in the given rectangular window and are in the universe of discourse [0, 1]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The universe of discourse is associate with fuzzy linguistic variable of feature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vertical position and horizontal position.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLOBAL FEATURES [43]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Describes the character as a whole. These shape features represent global</td>
<td>96.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>approximation. This method relies on the features e.g. Area, compactness. Some of</td>
<td>NOT MENTIONED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the global features are:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a) Start point and end point</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) Horizontal and vertical motion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) Pen-ups and number of sub-patterns</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEOMETRICAL FEATURES [43]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>It refers to sub-patterns or segments and give us a structured description of the</td>
<td>96.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>character or symbol as a loss of known features. Two classes are defined:</td>
<td>NOT MENTIONED</td>
<td></td>
<td></td>
</tr>
<tr>
<td>straight line and curved line.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHADOW FEATURES OF CHARACTER [44]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The rectangular boundary enclosing the character image is divided into eight</td>
<td>80.59%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>octants, for each octant shadow of character segment is computed on two perpendicular</td>
<td>MLP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sides, so total of 16 shadow features are obtained</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHAIN CODE HISTOGRAM CHARACTER CONTOUR [44]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Each pixel of the contour is assigned a different code that indicates the</td>
<td>64.90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>direction of the next pixel that belongs to the contour in some given direction.</td>
<td>MLP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTERSECTION FEATURE BASED [44]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intersection points are unique for a character in different segment. For</td>
<td>36.71%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>example, if image size is 25 x 25 pixels, the no. of features extracted are 32,</td>
<td>MLP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>out of which first 16 features represents number of open ends and rest represents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number of junction points within a segment.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BARR FEATURES [3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In this method four image parameters are generated, and every image parameter</td>
<td>91.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>corresponds to one the directions: east, north, and northeast, northwest. Every</td>
<td>MLP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>image parameter has a whole value representing the Barr length in a direction.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusions**

Postal automation is a topic of research interest for last two decades and many pieces of published article are available towards postal automation of non-Indian language documents [38, 41, 42]. At present postal sorting machines are available in several countries like USA, UK, Canada, Japan, France, Germany etc., but no postal automation machine is available for India. System development towards postal automation for a country like India is more difficult than other countries because of its multi-lingual and multi-script behavior. Although there are many languages and scripts in India, only a few pieces of work have been done towards postal automation in India [39, 40]. The work till on date is done on structured envelopes. No attempt has been made on non-structured envelopes which are major means of business transactions. For doing the Indian Postal system automation following things are required:-

1. Language Independent recognition structure.
2. Rapid and sustained adaptation to writers with different writing characteristics.
3. Immunity to erroneous or incomplete inputs.
4. The recognition of natural, unconstrained handwritten texts for handwritten words, and even for machine printed text, as many handwritten texts might, in real life tasks, be preceded or followed by machine printed data.
5. The development of language models specific to handwriting (punctuation marks, colon, semicolon etc.)
6. Sound measures on image quality so as to increase the recognition rate by improving the image acquisition process.
7. Rotation and scale invariant character recognition system. The existing methods are prone to various deformations, changes in size or shifts in the positions of characters. Existing techniques requires some sort of modifications for using them in postal system automation.
8. Need to re-examine, formulate and propose new algorithms for the classification of characters.
9. Need to improve the current document segmentation algorithms.
10. Processing of the document irrespective of its size, color etc.

**References**

man, and cybernetics-Part C: Applications and reviews, 31(2).


