# Offline Recognition of Image for content Based Retrieval

Smeet D. Thakur
*Department of Information Technology,PRMIT & R*

Prof. Smita S. Sikchi
*Department of Information Technology,PRMIT & R*

**Abstract— This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an IEEE conference. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.**

***Keywords*— Pre-processing, 2. Segmentation. 3. Feature Extraction, 4.Recognition, 5. Classification**

## I. INTRODUCTION

OCR work on printed Devnagari script started in early 1970s. Earlier studies on Devnagari script presented a Devnagari hand-printed numeral recognition system based on binary decision tree classifier. The study investigates the direction of the Devnagari Optical Character Recognition research (DOCR), analyzing the limitations of methodologies for the systems which can be classified based upon two major criteria: the data acquisition process (on-line or off-line) and the text type (machine-printed or hand-written). No matter which class the problem belongs, in general there are five major stages in the DOCR problem: 1. Pre-processing, 2. Segmentation. 3. Feature Extraction, 4.Recognition, 5. Classification The off-line and on-line character recognition techniques have different approaches; they share a lot of common problems and solutions. Since it is relatively more complex and requires more research compared to on-line and machine-printed recognition, off-line handwritten character recognition is selected as a focus of attention Handwriting Recognition Technology has been improving much under the purview of pattern recognition and image processing since a few decades. Hence various soft computing methods involved in other types of pattern and image recognition can as well be used for DOCR. Optical Character Recognition is a process by which we convert printed document or scanned page to ASCII character that a computer can recognize. The document image itself can be either machine printed or handwritten, or the combination of two. Computer system equipped with such an OCR system can improve the speed of input operation and decrease some possible human errors. Recognition of printed characters is itself a challenging problem since there is a variation of the same character due to change of fonts or introduction of different types of noises. Most of the Indian scripts are composed in two dimensions that make them different from Roman script. Therefore, the algorithms developed for Roman script are not directly applicable to Indian scripts. Many works on Indian scripts OCR have been reported. However, none of these works have considered real-life printed text in Devanagari. Consisting of character fusions and noisy environment. The present a complete OCR for printed text that is written in Devanagari script. The OCR has been tested on samples from various magazines and newspapers.

## II. DEVANAGARI OPTICAL CHARACTER RECOGNITION

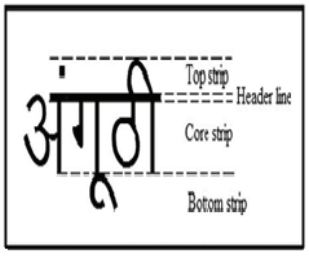
Most of the Indian scripts including Devanagari originated from ancient Brahmi script through various transformations. The script has a complex composition of its constituent symbols. Devanagari has 13 vowels and 34 consonants along with 14 modifiers of vowels and of “rakar,” Apart from the vowels and consonants, there are compound (composite) characters in most of Indian scripts including Devanagari, which are formed by combining two or more basic characters. The shape of a compound (composite) character is usually more complex than its constituent characters. A vowel following a consonant may take a modified shape, which depending on the vowel is placed to the left, right, top, or bottom of the consonant, and are called modifiers or “matras.” Text, characters, and digits are written from left to right in Devanagari. There is no concept of upper or lowercase characters [1]. It is a phonetic and syllabic script. As Devanagari is phonetic, words are written exactly as they are pronounced; syllabic means that text is written using consonants and vowels that together form syllables. The vowels in can be the vowels in can be either independent or dependent. The script uses modifiers for “nasalization” or aspiration of a vowel or a consonant. Every Indian script has its own specified

composition rules for combining vowels, consonants, and modifiers. Some of them can be combined with their type. A modifier can be attached to a vowel or to a consonant. Consonants may have a half form when they are combined with other consonants as depicted in Except for some characters, the half forms of consonants are the left part of original consonants with the right part removed. Some special combinations are also where a new character or the half forms of consonants may appear in the lower half of the new composite forms. Another distinctive feature of Devanagari is the presence of a horizontal line on the top of all characters. This line is known as header line or "shirorekha". The words can typically be divided into three strips: top, core, and bottom. The header line separates the top and core strips and a virtual base line separates the core and lower strips. The top strip generally contains the top modifiers, and bottom strip contains lower modifiers. When two or more characters appear side by side to form a word in Devanagari, the header lines touch and generate a bigger header line. To get an idea about the occurrence frequency of different Devanagari characters, the occurrence statistics of 20 frequent characters in Devanagari script, based on a study of three million words. They also noted from a statistical analysis that about 99% of Devanagari words have header lines. As far as the development of an OCR system is concerned, the most important step can be the extraction of precise features from the characters, symbols, and words. There are several ways for feature extraction, but the most important is to extract the features, which can distinct different patterns. Al-most all the features can be of two types, namely statistical and structural [6]. The statistical features are derived from statistical distributions of points, such as moments, zoning, histograms, or projection features. Structural features are mainly based on geometrical properties of a symbol (character), like loops, directions of strokes, intersections of strokes, and end points. Since statistical and structural features complement each other, many researchers have tried to integrate them to highlight different properties of a pattern. The recognition or classification process of characters, symbols, or words is normally carried out using template or feature-based approaches. In the template-based approach, an unknown test pattern is compared directly with the ideal pattern and the degree of correlation between the two is used for classification or taking a decision. Feature-based approaches extract features from the test patterns and use them in classification models like artificial neural networks hidden Markov models support vector machines modified quadratic discriminate function etc. Some contemporary techniques like rough sets, fuzzy rules, evolutionary algorithms, pdist are also used for the recognition purpose of Devanagari characters. The automatic processing of a document image containing handwritten or machine-printed information can be divided into three stages region extraction and categorization using document layout analysis; text line, word, and character segmentation; and individual character recognition[1].


Fig. 1 (a) Vowels and modifiers of Devanagari script. (b) Consonants and their corresponding half forms in Devanagari (c) Some combinations of consonants with themselves.

| Symbol | Occurance(%) | Symbol | Occurance(%) |
|---|---|---|---|
| ा | 10.126 | ं | 3.754 |
| क | 7.949 | म | 3.360 |
| र | 7.409 | ो | 2.911 |
| ि | 6.532 | य | 2.771 |
| न | 5.064 | प | 2.473 |
| ी | 4.478 | ल | 2.456 |
| ह | 4.396 | व | 2.268 |
| स | 4.368 | द | 1.720 |
| त | 4.324 | ज | 1.462 |
| ि | 4.224 | ग | 1.334 |

(a)




(b)

Fig. 2 (a) Occurrence statistics of 20 frequent characters In Devnagari.
(b) Three strips of a word in Devanagari script

## III. OBJECTIVE & SCOPE

Research and development in Indian language processing is a necessity for a highly multilingual, multiple-script country like India. Ministry of Information Technology of Government of India started a program on Technology Development for Indian Languages where language aspects are studied and developed. Another Government undertaking CDAC (Centre for Development of Advance Computing) is actively involved in development of Indian languages fonts, translators). Various hardware and software based language processors and language translators are developed by CDAC in collaboration with IIT Kanpur and indigenously (GIST, LIPI, ISM for word processing and Chitrankan software for offline character recognition ISCII (Indian Scripts Standard Code for Information Interchange), the Indian standards for various languages was developed in 1988 by Indian Government Also various Vedic script symbols are incorporated in Unicode consortium. Researchers have investigated OCR for a number of Indian scripts: Devnagari, Tamil, Telugu, Bengali, and Kannada, Gurumukhi. However, most of this research has been confined to the identification of isolated characters rather than the script. Some systems used a statistical method; others were syntactic and/or heuristic-based. Unlike roman script, the Indic scripts are a composition of the constituent symbols in two dimensions. In conventional Research, first a word is segmented into its composite characters. Each composite character is then decomposed into the constituent symbols or the strokes (diacritic marks like matra) that are finally recognized. Holistic approaches circumvent the issues of segmentation ambiguity and character shape variability that are primary concerns for analytical approaches, and they may succeed on poorly written words where analytical methods fail to identify character content. A lot of research is still needed for word, sentence and document recognition, its semantics and lexicon. There is still a dearth of need to do the research in the area Devnagari character recognition.

## IV. LITERATURE REVIEW

India is a multi-lingual and multi-script country comprising of eighteen official languages. One of the defining aspects of Indian script is the repertoire of sounds it has to support. Because there is typically a letter for each of the phonemes in Indian languages, the alphabet set tends to be quite large. Most of the Indian languages originated from Bramhi script. These scripts are used for two distinct major linguistic groups, Indo-European languages in the north, and Dravidian languages in the south. Devnagari is the most popular script in India. It has vowels and consonants[9]. They are called basic characters. Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. When vowels are written in this way they are known as modifiers and the characters so formed are called conjuncts. Sometimes two or more consonants can combine and take new shapes. These new shape clusters are known as compound characters. These types of basic characters, compound characters and modifiers are present not only in Devnagari but also in other scripts. Hindi, the national language of India, is written in the Devnagari script. Devnagari is also used for writing Marathi, Sanskrit and Nepali. Moreover, Hindi is the third most popular language in the world Various approaches used for the design of DOCR systems are discussed below:

i) Matrix Matching: Matrix Matching converts each character into a pattern within a matrix, and then compares the pattern with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.

ii) Structural Analysis: Structural Analysis identifies characters by examining their sub features- shape of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints.

iii) Neural Networks: This strategy simulates the way the human neural system works. It samples the pixels in each image and matches them to a known index of character pixel patterns. The ability to recognize characters through abstraction is great for faxed documents and damaged text. Neural networks are ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns. Neural networks is two complimentary technologies neural networks can learn from data and feedback [4]. It is difficult to develop an insight about the meaning associated with each neuron and each weight. Viewed as “black box” approach (know what the box does but not how it is done conceptually!)

## V. PROPOSED SYSTEM

1. Pre-processing
2. Segmentation
3. Feature Extraction

4. Recognition
5. Classification

*1. Pre-processing*

In preprocessing, select database from created database. Croping of database of consonants and modifiers. Then it detects the maximum area and calculate centroid red mark star.

Data in a paper document are usually captured by optical scanning and stored in a file of lecture elements, called pixels. These pixels may have values: OFF (0) or ON (1) for binary images, 0– 255 for gray-scale images, and 3 channels of 0–255 colour values for color images. This collected raw data must be further analyzed to get useful information. Such processing includes the following:

i) Thresholding: A grayscale or color image is reduced to a binary image.

ii) Noise reduction: 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

iii) Skew Detection and Correction: Handwritten document may originally be skewed or skewness may introduce in document scanning process. This effect is unintentional in many real cases, and it should be eliminated because it dramatically reduces the accuracy of the subsequent processes, such as segmentation and classification. Skewed lines are made horizontal by calculating skew angle and making proper correction in the raw image

iv) Size Normalization: Each segmented character is normalized to fit within suitable matrix like 32x32 or 64x64 so that all characters have same data size.

v) Thinning: The boundary detection of image is done to enable easier subsequent detection of pertinent features and objects of interest.


Fig. 3 Preprocessed Images (a) Original (b) segmented
(c) Shirorekha removed (d) Thinned (e) image edging

*2. Segmentation*

It collects data from different source. Segmentation is one of the most important phases of DOCR system. By applying good segmentation techniques we can increase the performance of DOCR. Segmentation subdivides an image into its constituent regions or objects. Basically in segmentation, we try to extract basic constituent of the script, which are certainly characters. This is needed because our classifier recognizes these characters only. Segmentation phase is also crucial in contributing to this error due to touching characters, which the classifier cannot properly tackle. Even in good quality documents, some adjacent characters touch each other due to inappropriate scanning resolution. Numbers of constituent characters touching each other in Devnagari scripts are shown in table 1. To tackle the touching characters in Devnagari documents, at first, we attempt to identify the touching characters. Next, they are segmented into constituent ones using a fuzzy decision making approach. In Devnagari script, a text word may be partitioned into three zones. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic and compound characters below the headline, and the lower zone may contain where some vowel and consonant modifiers can reside. For a long number of characters (basic as well as compound) there exists a horizontal line at the upper part called “shirorekha” or headline in Hindi. The imaginary line separating the middle and lower zone may be called the base line. Line, Word and Character Segmentation: Once the text blocks are detected, the OCR system automatically finds individual text lines, segments the words, and then separates the characters accurately. Segmentation of Line: Text lines are detected by horizontal scanning. For segmentation of line, we

scan scanned document page horizontally from the top and find the last row containing all white pixels, before a black pixel is found. Then we find the first row containing entire white pixel just after the end of black pixels. We repeated this process on entire page to find out all lines. Segmentation of Words: After finding a particular line we separate individual words. This is done by vertical scanning. Segmentation of Individual Characters: Once we get the words we segment it to individual characters. Before segmenting words to individual characters, we locate the head line. This is done by finding the rows having maximum number of black pixels in a word. After locating head line we remove it i.e. converts it in white pixels. After removing head line our word is divided into three horizontal parts known as upper zone, middle zone and lower zone. Individual characters are separated from each zone by applying vertical scanning in Devnagari script; a text word may be partitioned into three zones. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic and compound characters below the headline, and the lower zone may contain where some vowel and consonant modifiers can reside. For a long number of characters (basic as well as compound) there exists a horizontal line at the upper part called "shirorekha" or headline in Hindi. The imaginary line separating the middle and lower zone may be called the base line.

*3. Feature Extraction*

Feature extraction and selection can be defined as extracting the most representative information from the raw data, which minimizes the within class pattern variability while enhancing the between class pattern variability. For this purpose, a set of features are extracted for each class that helps distinguish it from other classes, while remaining invariant to characteristic differences within the class. Feature is a point of human interest in an image, a place where something happens. It could be an intersection between two lines, or it could be a corner open end or it could be just a dot surrounded by space. These relationships are used for character identification. Intersection features are unique for different characters; hence the feature points are exploited for the task of character recognition. Each handwritten character can be adequately represented within 16 segments (each of size 25 X 25 pixels) and hence 32 features for each character can be used as input to neural network. We are using a discrete structural approach and breaking the character boundary into 16 segments. For each segment number of intersection points, number of open ends is being calculated. It divides images in 8 to 16 parts.

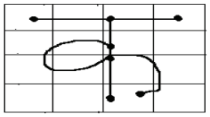


Fig. 4 Intersection points of character

Fig. Intersection point of character

GLCM (grey level occurrence matrix), Affine, Colored domino, histogram are the feature of images in feature extraction.

Affine Moment Invariant (AMIs): The AMIs were derived by means of the theory of algebraic invariants. The AMIs is invariant under general affine transformation

$$u = a_0 + a_1 x + a_2 y \quad \text{——————} \quad 1$$

$$v = b_0 + b_1 x + b_2 y \quad \text{——————} \quad 2$$

*3.1 Reduction*

Fig. 5 Reduction of Element

Thinning is the process to extract and apply additional constraints on the pixel elements that need to be preserved such that a linear structure of the input image will be recaptured without destroying its connectivity. In the context of image

*3.2 Box Method*

The feature of a given character is extracted in following steps

1) A given number is divided in equal blocks having (5x5) =25 pixels in each block

2) A weight value is assigned to each block. It is equal to number of pixels in the block. One example is given

3) The weight function is normalized, with respect to 100.
4) The values of the weight functions along with its (i,j) th

*4. Recognition*

Gives as an input image to recognize. The process of conversion of scanned image into a text document primarily consists of the following steps:

- Preprocessing: The preprocessing steps remove any distortions or discontinuity in the input character and convert the characters into a form recognizable by the detection procedure. It consists of following steps:

1. Size Determination: This step determines the approximate dimension of the character by forming a tight fit rectangular boundary around the character.
2. Distortion Removal: We use thickening, thinning and pruning for removing distortions. The image is thickened first and then thinned to convergence. This gives us a smooth one-pixel wide image of the character, which is pruned to remove the small projections resulting from the thinning algorithm. Small characters should be distortion free.
3. Normalization: After thinning character is scaled to 100 X 100 pixels using affine transformation.

*5. Classification*

Before applying Neural network, a prelimimary classification is performed for better results. The presence and position of spine divides character set of devanagri into different classes, so the entire character map for Devanagari characters (excluding matras) can be grouped according to the following criteria

*1. Shirorekha Continuity:*
Some characters contain a shirorekha throughout, while the others contain a partial shirorekha or no shirorekha. Thus three groups can be obtained by this method.

*2. Spine Location:*
Another important aspect of the character is its "spine". Characters can be divided into three groups
i) End Spine (The spine is the rightmost part of the character)
ii) Mid Spine (The spine exists in between, i.e. there are some parts of the character on either side of the spine)
iii) No Spine (There are no spines in these characters). By taking an intersection of the above two properties, the entire character map can be divided into small groups. This eases the task of recognition.

As a classifier ANFIS (Advance Neuro Fuzzy Logic System) for neural network and fuzzy logic. Neural networks and fuzzy logic are two complimentary technologies.

Pdist Technique for all i.e. for segmentation, feature extraction, recognition, and classification.

Pair wise distance between pairs of objects

*Syntax*

D=pdist(x)
D = pdist(X,*distance*)

*Description*

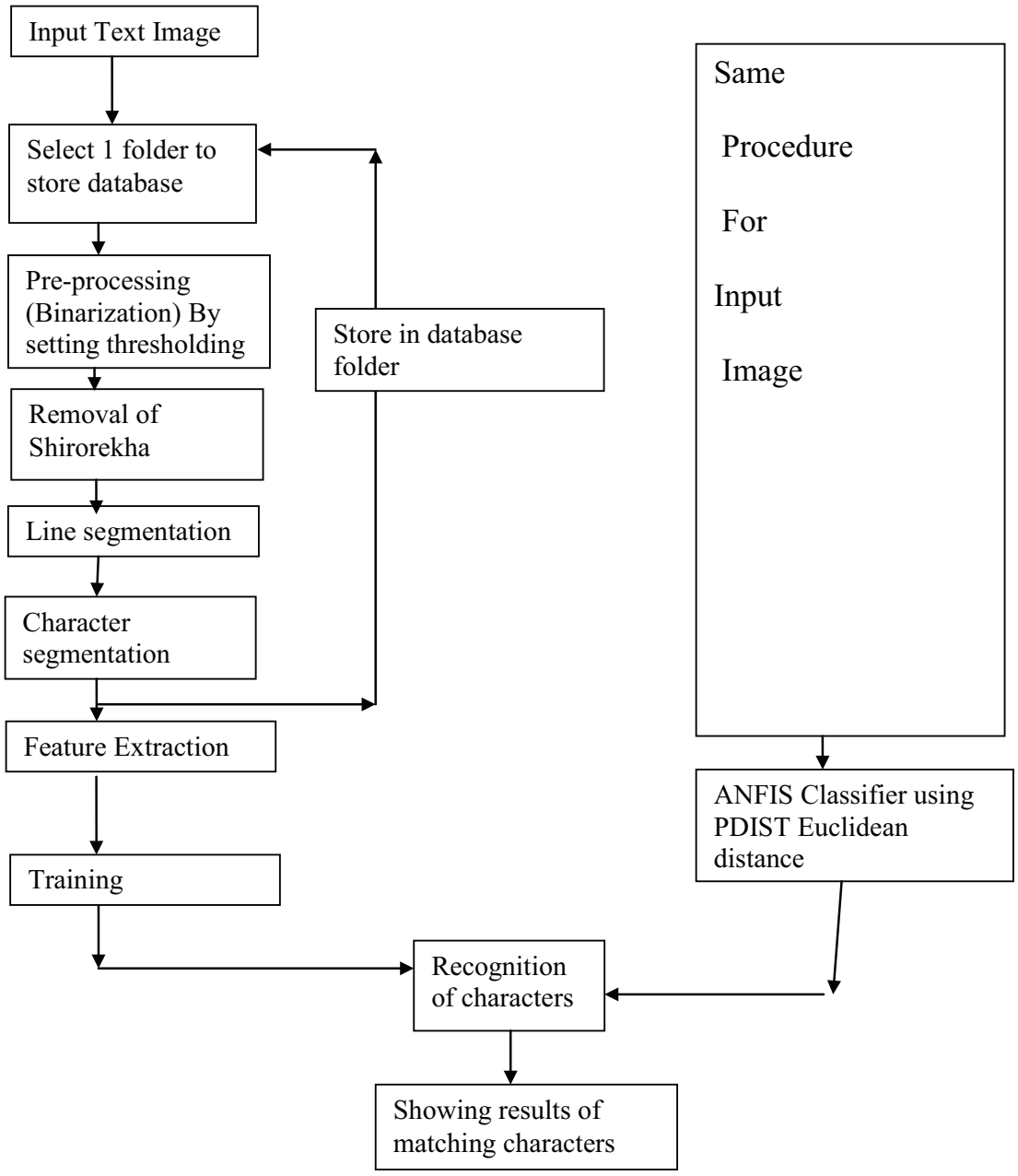
D = pdist(X) computes the Euclidean distance between pairs of objects in *m*-by-*n* data matrix X. Rows of X correspond to observations, and columns correspond to variables. D is a row vector of length *m*(*m*–1)/2, corresponding to pairs of observations in X. The distances are arranged in the order (2,1), (3,1), ..., (*m*,1), (3,2), ..., (*m*,2), ..., (*m*,*m*–1)). D is commonly used as a dissimilarity matrix in clustering or multidimensional scaling.

To save space and computation time, D is formatted as a vector. However, you can convert this vector into a square matrix using the square form function so that element $i, j$ in the matrix, where $i < j$, corresponds to the distance between objects $i$ and $j$ in the original data set.

D = pdist (X, *distance*) computes the distance between objects in the data matrix, X, using the method specified by *distance*, which can be any of the following character strings.

## VI. SYSTEM DESIGN

*VI.1 Flow Chart*



*VI.2 System design/ Block diag.*

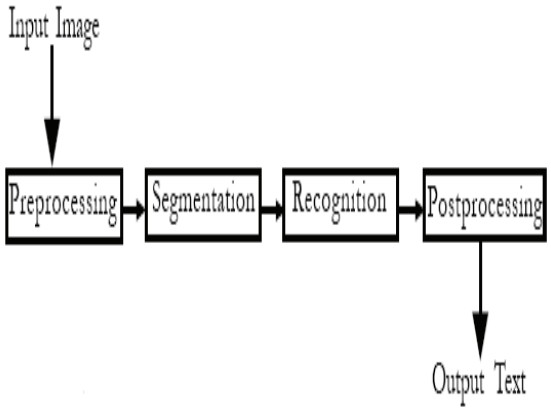



Fig. 6 Block Diagram of OCR

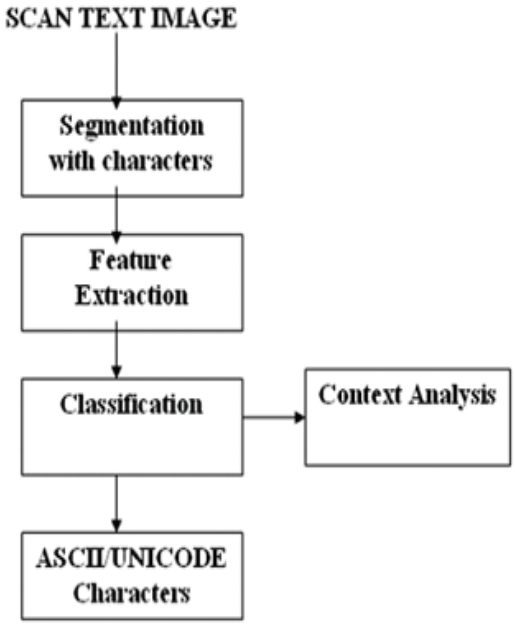



Fig. 7 Stages in OCR Design

*VI.3 System Requirement*

Tool : MATLAB
Languages : C++, Java, C#
O/S : Windows XP (Professional)
Database : My SQL/ Oracle 9i

## VII. CONCLUSION

The With the advent of computer and information technology, there has been a dramatic increase of research in the field of Devanagari OCR since 1990. Different strategies using combination of multiple features,

multiple classifiers, and multiple templates have been considered extensively in the state of the art. Only a few works have been reported in the areas of un-constraint Devanagari handwriting recognition. Lexicon-based approaches shall be used for recognizing legal amounts on bank cheques and city names on postal documents. In countries like India, where many languages and scripts exist, the identification of script has to be done prior to the recognition in applications like postal address reader, where address can be written in any Indian script. In India huge volumes of historical documents and books (handwritten or printed in Devanagari script) remain to be digitized for better access, sharing, indexing, etc. This will definitely be helpful for other research communities in India in the areas of social sciences, economics, and linguistics. The errors in recognizing printed Devanagari characters are mainly due to incorrect character segmentation of touching or broken characters. Because of upper and lower modifiers of Devanagari text, many portions of two consecutive lines may also overlap and proper segmentation of such overlapped portions are needed to get higher accuracy.

## REFERENCES


[1] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal "Offline Recognition of Devanagari Script: A Survey" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 41, NO. 6, NOVEMBER 2011

[2] Vikas J Dongre Vijay H Mankar "A Review of Research on Devnagari Character Recognition" International Journal of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010

[3] Raghuraj Singh, C. S. Yadav, Prabhat Verma, Vibhash Yadav "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network" International Journal of Computer Science & Communication Vol. 1, No. 1, January-June 2010, pp. 91-95

[4] Sandhya Arora, Meghnad Saha, Debotosh Bhattacharjee, Mita Nasipuri, Latesh Malik "A Two Stage Classification Approach for Handwritten Devanagari Characters"

[5] M. Hanmandlu and Pooja Agrawal "A Structural Approach for Segmentation of Handwritten Hindi Text" Proceedings of the International Conference on Cognition and Recognition.

[6] Nafiz Arica and Fatos T. Yarman-Vural "An Overview of Character Recognition Focused on Off-Line Handwriting" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 31, NO. 2, MAY 2001

[7] Holambe A.N. , Thool R.C., Shinde U.B. and Holambe S.N. "Brief review of research on Devanagari script" International Journal of Computational Intelligence Techniques, ISSN: 0976–0466, Volume 1, Issue 2, 2010, pp-06-09

[8] M. Meshesha and C. V. Jawahar, "Matching word images for contentbased retrieval from printed document images," Int. J. Document Anal. Recognit., vol. 11, pp. 29–38, 2008

[9] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal "Offline Recognition of Devanagari Script: A Survey" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS PART C: APPLICATIONS AND REVIEWS, VOL. 41, NO. 6, NOVEMBER 2011

[10] Veena Bansal and R.M.K. Sinha, On how to describe shapes of Devanagari characters and use them for recognition, in Proceedings - Fifth International Conference on Document Analysis and Recognition IEEE Publication, held at Bangalore from Sep 21- 23, 1999, pp. 653-656.

[11] R. M. K. Sinha, "A journey from Indian scripts processing to Indian language processing," IEEE Ann. Hist. Comput., vol. 31, no. 1, pp. 8– 31, Jan./Mar. 2009.