**A REVIEW PAPER ON:DOCUMENT IMAGE ANALYSIS FOR HANDWRITTEN DEVNAGRI SCRIPT**

**HARSHA JAIN**

**DEPARTEMENT OF COMPUTER SCIENCE AND ENGINNERING,AMITY UNIVERSITY(M.P)**

**Abstract**—Handwritten Devanagari script recognition system using neural network is presented in this paper. Diagonal based feature extraction is used for extracting features of the handwritten Devanagari script. After that these feature of each character image is converted into chromosome bit string of length 378. More than 1000 sample is used for training and testing purpose in this proposed work. It is attempted to use the power of genetic algorithm to recognize the character. In step-I preprocessing on the character image, then image suitable for feature extraction as here is used. Diagonal based feature extraction method to extract 54 features to each character. In the next step character recognize image in which extracted feature in converted into Chromosome bit string of size 378. In recognition step using fitness function in which find the Chromosome difference between unknown character and Chromosome which are store in data base. There has been a significant amount of research in various aspects of writing based user interfaces including interactive design tools, studies of gestures, software toolkits, ink beautification, and sketch recognition. In this paper, we shall focus on the recognition of handwritten characters that are used in common applications. Different people may use different stroke-order, number, and direction to draw the same shape of any character. Handwritten characters are imprecise in nature such that corners are not always sharp, lines are not perfectly straight, and curves are not necessarily smooth. A robust recognition system has to account for all of these factors. In this paper, we shall consider a statistical approach to Handwritten Character Recognition using Zernike moments as features. The rotation invariance property of Zernike moments is especially desirable for Handwritten Character Recognition.

**Keywords Classification** Devnagari Character Recognition, Handwriting Recognition,Segmentation,

Feature Extraction.

**Introduction**

In the late 1980’s, the prevalence of fast computers, large computer memory, and inexpensive scanners fostered an increasing interest in document image analysis. With many paper documents being sent and received via fax machines and being stored digitally in large document databases, the interest grew to do more with these images than simply view and print them. The objective of document image analysis is to recognize the text and graphics components in

images, and to extract the intended information as a human would. Two categories of document

image analysis can be defined (see Figure 1). Textual processing deals with the text components

of a document image. Some tasks here are: recognizing the text by optical character recognition

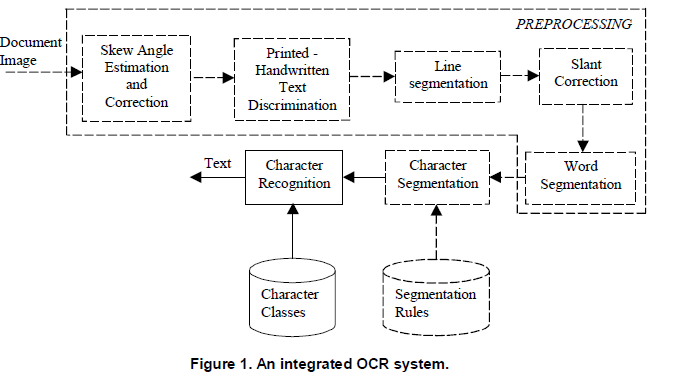
(OCR), determining the skew (any tilt at which the document may have been scanned into the

computer), finding columns, paragraphs, text lines, and words. Graphics processing deals with the

non-textual line and symbol components that make up line diagrams, delimiting straight lines

between text sections, company logos, etc. Because many of these graphics components consist of

lines, such processing includes line thinning, line fitting, and corner and curve detection



Handwriting recognition has been one of the fascinating and challenging research areas in field of image processing and pattern recognition in the recent years**.**O NE interesting and challenging ﬁeld of research in pattern recognition is Optical Character Recognition (OCR). Optical character recognition is the process in which a paper document is optically scanned and then converted into computer processable electronic format by recognizing and associating symbolic identity with every individual character in the document. With the increasing demand for creating a paperless world, many OCR algorithms have been developed over the years . However, most OCR systems are script-speciﬁc in the sense that they can read characters written in one particular script only. Script is deﬁned as the graphic form of the writing system used to write statements expressible in language. That means, a script class refers to a particular style of writing and the set of characters used in it. In general, handwriting recognition classified into two types as off-line and on-line handwriting recognition methods. In off-line recognition, the writing is usually captured optically by a scanner and complete writing is available as an image. But, in the on-line system the two dimensional coordinates of successive points are represented as a function of time and the order of strokes made by the writer are also available. The on-line methods have been shown to be superior to their off-line counterparts in recognizing handwritten characters due. to the temporal information available with the former . Several applications including mail sorting, bank processing, document reading and postal address recognition require off-line handwriting recognition systems. As the result, the off-line handwriting recognition continues to be an active area for research towards exploring the newer techniques that would improve recognition accuracy. 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. Post processing**

**Pre-processing:**

Pre-processing followed by segmentation and feature extraction. Pre-processing includes the steps that are required to shape the input image into a form suitable for segmentation. The pre-processing is a series of operation performed on the scanner input image. It essentially enhances the image rendering it suitable for segmentation. The various tasks performed on the image in pre-processing stage shown in Fig. 2. Binarization process converts a gray scale image into a binary image using thresh holding technique. Detection of edges in the binarized image using canny technique, dilation the image are the operations performed in the last two stages to produce the pre-processed image suitable for segmentation.

Scanned input Image



Noise Removing



Binarization



Edge Detection



Dilation and Filling



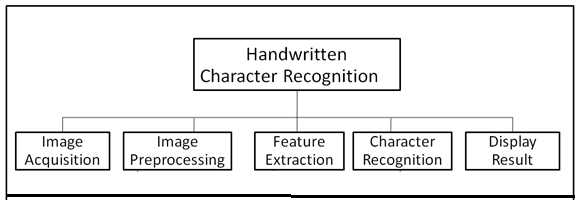
Processed Image for feature

Extraction

**Fig: Pre-processing of handwritten character**

**Segmentation**

In the segmentation stage, an image of sequence of characters is decomposed into sub-images of individual character. In the proposed system, the pre-processed input image is segmented into isolated characters by assigning a number to each character using a labeling process. This labeling provides information about number of characters in the image. Each individual character is uniformly resized into 90x60 pixels for classification and recognition stage.



**Feature Extraction Method** 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 [25]. A good survey on feature extraction methods for character recognition can be found in [26]. Various feature extraction methods are classified in three major groups:

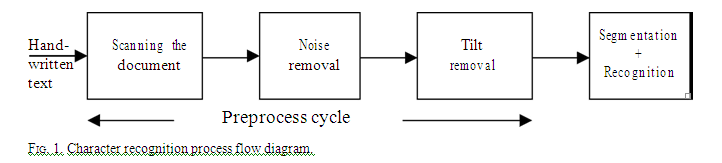
1. Global Transformation and Series Expansion

2. Statistical Features

3. Geometrical and Topological Features

**Classification and Recognition**

The final step is the recognition and description step, where components are assigned a semantic label and the entire document is described as a whole. It is at this stage that domain knowledge is used most extensively. The result is a description of a document as a human would give it. For a text image, we refer not to pixel groups, or blobs of black on white, but to titles, subtitles, bodies of text, footnotes, etc. Depending on the arrangement of these text blocks, a page of text may be a title page of a paper, a table of contents of a journal, a business form, or the face of a mail piece. For a graphical image, an electrical circuit diagram for instance, we refer not to lines joining circles and triangles and other shapes, but to connections between AND gates, transistors and other electronic components. The components and their connections describe a particular circuit that has a purpose in the known domain. It is this semantic description that is most efficiently stored and most effectively used for common tasks such as indexing and modifying particular document components



**FEATURES OF DEVNAGARI SCRIPT**

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 11 vowels and 33 consonants. 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.

**IMAGE PREPROCESSING**

Data in a paper document are usually captured by optical scanning and stored in a file of picture 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 colour images. This collected raw data must be further analyzed to get useful information. Such processing includes the following:

**Thresholding:**

A grayscale or colour image is reduced to a binary image.

**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.

**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.

**Size Normalization:**

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

**Thinning:**

The boundary detection of image is done to enable easier subsequent detection of pertinent features and objects of interest . Various standard functions are now available in MATLAB for above operations.

**FUTURE RESEARCH**

Research and development in India 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 (TDIL: http://www.tdil.mit.gov.in) 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). R M K Sinha of IIT Kanpur has been instrumental in the development of Indian language recognition and processing since the beginning. ISCII (Indian Scripts Standard Code for Information Interchange), the Indian standards for various languages was developed in 1988 by Indian Government. Also various Indic script symbols are incorporated in Unicode consortium (http://www.unicode.org/ charts/PDF/U0980.pdf). Research in Devnagari character is currently undergoing in leading institutes in IIT Kanpur, IIIT Hydrabad, ISI Kolkata and many others.

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.

**CONCLUSION**

Methods for treating the problem of Devnagari character recognition have developed remarkably in the last two decades. Still a lot of research is needed to tackle the challenges in DOCR so that commercially viable software solutions can be made available. It is hoped that this comprehensive discussion will provide insight into various concepts involved, and boost further advances in the area. The difficulty of performing accurate recognition is determined by the nature of the material to be read and by its quality. G handwritten writing. Methods of increasing sophistication are being pursued. Current research employs models not only of characters, but also words and phrases, and even entire documents. The powerful tools such as HMM, neural networks and their combinations are used. In order to have high reliability in character recognition, segmentation and classification have to be treated in an integrated manner to Generally, misrecognition rates for unconstrained material increase progressively from machine print to

obtain more accuracy in complex cases. This paper has concentrated on an appreciation of principles and methods. Present work has not attempted to compare the effectiveness of various algorithms. It would be difficult to assess techniques separate from the systems for which they were developed. Unfortunately there is little experimental as well as standard handwritten character database available publicly for The list of references to provide more detailed understanding of the approaches described is enlisted. We apologize to researchers whose important contributions may have been overlooked. Benchmarking the accuracy of various advanced techniques proposed in Devnagari character recognition .

**REFERENCES [1]** B.B.Chaudhuri and U.Pal,” An OCR system to read two

Indian language scripts: Bangla and Devnagari (Hindi)”,

*Proc. of 4th ICDAR*, Uhn. 18-20 August 1997.

**[2]** B.B.Chaudhuri and U.Pal, “A complete printed Bangla

OCR,” *Pattern Recognition*, vol. 31, pp. 531-549, 1998.

**[3]** Santanu Chaudhury, Gaurav Harit, Shekar Madnani,

R.B.Shet,” Identification of scripts of Indian languages by

Combining trainable classifiers,” *Proc. of ICVGIP*, Dec-20-

22, Bangalore, India, 2000.

**[4]**

D.Dhanya, A.G Ramakrishnan and Peeta Basa pati, “Script

identification in printed bilingual documents,” *Sadhana,* vol.

27, part-1, pp. 73-82, 2002.

**[5]**

R. Sanjeev Kunte1 and R.D. Sudhaker Samuel, **“**On

Separation of Kannada and English Words from a Bilingual

Document Employing Gabor Features and Radial Basis

Function Neural Network”, *Proc. of ICCR,* pp. 640-644,

2005.

**[6]** M.C.Padma and P. Nagabhushan,” Script Identification and

separation of text words of Kannada Hindi and English

languages through discriminating features,” *Proc. of*

*NCDAR-2003*, pp. 252-260. 2003.

**[7]** G.S.Peake and Tan, “Script and language identification from

document images,” *Proc. of Eighth British Mach. Vision*

*Conf*., vol.2, pp. 230-233, Sept-1997.

[10] U.Pal and B.B.Chaudhuri, “Script line separation from

Indian Multi-script documents,” *5th ICDAR,* pp.406-409,

1999