Chapter 1

Introduction

1.1 Optical Character Recognition (OCR)
   1.1.1 Types of OCR
   1.1.2 Data collection
   1.1.3 Pre-processing
   1.1.4 Segmentation
   1.1.5 Feature extraction
   1.1.6 Classification

1.2 Literature Review
1.3 Motivation for the present work, Problem Statement
1.4 Organization of Thesis

The thesis entitled 'A Study on Handwritten Marathi Word Recognition' presented here is OCR for handwritten Marathi words. OCR is acronym for Optical Character Recognition in which text images are converted into digital text without human intervention. This technology converts read only documents into digitized formats that can easily be retrieved, searched, and archived. Document analysis and recognition are two challenging research areas in pattern recognition. Although sufficient amount of research work is reported for printed offline OCR, little research work exists for offline handwritten OCR due to the diversified nature in handwritings. Handwritten Marathi word recognition is a challenging task because the total number of characters present in Marathi large. Also Marathi consists of various modifiers and different forms of compound characters which complicate the design of OCR procedures. In this chapter, we
Chapter 1: Introduction

give a brief description of OCR, literature review, motivation for the present work and problem statement.

1.1 Optical Character Recognition (OCR):

Optical Character Recognition (OCR) converts text images into digital text without user intervention. Since OCR has numerous applications like postal automation, automatic form processing, historical document preservation and many more, OCR is an area of interest for researchers working in document analysis and recognition. OCR can be broadly classified into two types: Online OCR and Offline OCR.

1.1.1 Types of OCR:

**Online OCR**: Online OCR converts input text to digital text as it is entered on the device. Device used for input text can be a mobile, PDA or any special digitizer. Pen movement, strokes and pen up and downs are recorded by the sensors which can be used for recognition purpose.

**Offline OCR**: Offline OCR converts printed/handwritten text images into digital text. Printed/handwritten texts documents are scanned using a scanner and converted into digital text so that computer understands and processes that text. The important steps in offline OCR are shown in the Figure 1.1.

Offline OCR can be broadly classified into two types: Printed OCR and Handwritten OCR.

**Printed OCR Vs Handwritten OCR**: Input for offline OCR is either printed documents or handwritten documents. Offline OCR converts printed/handwritten documents into digital text. Printed OCR is used to digitize historical documents,
books and printed forms. Handwritten OCR is used to digitize handwritten documents. Designing and developing handwritten OCR is more complicated and challenging task than printed OCR. Printed text has specific font type and specific size where as in handwritten text considerable variation exists as each person has different writing styles. Also segmentation process is difficult in handwritten OCR as compared to printed OCR. In handwritten OCR segmentation of text into characters is complicated task and which further reduces recognition accuracy. Major steps in offline OCR are discussed below:

1.1.2 Data Collection:

In order to develop offline handwritten OCR, database of handwritten sample images is to be created. Database has to be large in vocabulary and variations. There are standard databases such as CEDAR, NIST and CENPARMI which are used for experimentation of offline handwritten OCR. But all these databases are only for isolated English and Devanagari characters, but not for words, that is, strings of characters.
1.1.3 Pre-processing:

Pre-processing and image enhancement operations on images are carried out to improve the quality of image data and to remove distortions. We have to analyze information in the image so as to improve the quality and reduce distortions. First, image data is transformed to gray scale using Ostu’s threshold technique then gray scale image is converted to black and white image using binarization. Conversion from a gray-scale image to a black-and-white image may cause some loss of information. Dilation and erosion operations can reduce this loss of information. We can remove noise by using appropriate structuring elements. Slant corrections can be made to improve recognition rate. Also, we have to normalize images to a specified standard plane. Normalization is carried out to reduce the interclass variation of the shapes of the images. To carry out feature extraction and classification process, preprocessing and image enhancement is to be executed correctly; otherwise it may degrade the quality of image and important information may be lost.

1.1.4 Segmentation:

Segmentation divides an image into meaningful components called segments. Segmentation is of two types: contextual segmentation and non contextual segmentation. Contextual segmentation is more useful to differentiate objects using the pixels belongs to that object. There are two types of contextual segmentation depending on signal discontinuity and signal similarity. Cluster, compression based methods, histograms, edge detection are widely used in contextual segmentation. Non contextual segmentation differentiates the pixels irrespective of their location. A simple method of non contextual segmentation is thresholding. The accuracy of offline handwritten OCR recognition largely depends on the success of the segmentation phase.
1.1.5 **Feature extraction:**

Feature extraction is an important phase in OCR which occurs prior to classification. Recognition accuracy of OCR is largely depends on the extracted features. In this phase unique characteristics (features) are stored in a feature vector for all input images. Features are broadly classified into two types (i) global features, and (ii) geometric and topological features based on their characteristics:

**Global Features:**

Global features are also known as statistical features. Global features are not affected by noise or distortions and can be detected easily. Some commonly used global features are moments, zoning, projection, histogram, n-tuples, crossing and distances.

**Geometric and Topological Features:**

Geometric and topological features may represent both global and local properties, but are not affected by distortions or style variations. Object components, structure of objects and their properties can be represented using geometric and topological features. Geometric and topological representations can be broadly grouped into four categories:

i) Topological structures like lines, curves, loops, end points, branch points, T- point etc.

ii) Approximating geometric properties like aspect ratio, difference between x and y coordinates etc are a kind of representation.

iii) Codings like freeman chain code, normalized chain code, regular expressions are forms of another representation of geometric features.

iv) Graphs and trees are another type of representations, in which first topological features are extracted and those features are represented in graph or tree formats.
Chapter 1: Introduction

1.1.6 Classification:

Image classification assigns a label to an unknown object. Classification is broadly categorized into two types: supervised classification and unsupervised classification.

**Supervised classification:**

In supervised classification training data is used where available predefined class labels and features are used to assign labels to unknown objects. Supervised classification is appropriate when sufficient amount of training data is available.

**Unsupervised classification:**

Unsupervised classification is more appropriate when there is less information for classification. In unsupervised classification classes or groups are formed according to randomly sampled data called clusters and unknown objects are classified into these clusters.

Using various decision rules, unknown objects are classified into respective classes. Some commonly used classification techniques and decision rules are discussed below:

**Multilevel slice classifier:**

Multilevel slice classifier decision rule is defined on the basis of lowest and highest values of classes. This classifier is also known as parallelepiped classifier is very simple and easy to understand. In this case classification accuracy depends on the lowest and highest values of classes chosen.

**Minimum distance classifier:**

In minimum distance classifier unknown image is classified into a class that minimizes the distance between the image and the class. Decision rule is
Chapter 1: Introduction

based on the distances of image from the classes. Generally, Euclidian distance or Mahalanobis distance are used in minimum distance classifiers.

**Maximum likelihood classifier:**

Maximum likelihood classifier decision rule is based on the posterior probability of a pixel belonging to the class.

**Fuzzy set theory and expert system:**

Fuzzy set theory uses a ‘membership function’. It is difficult to define an appropriate membership function and boundaries of different classes for classification. Fuzzy set theory based classifiers are useful for qualitative data. Expert system classifiers use knowledge based on experiences.

For the present work support vector machine and k-NN classifiers are considered.

**1.2 Literature Review:**

Handwriting recognition is one of the important research problems in the field of document analysis and recognition. Document analysis and recognition is challenging area in pattern recognition due to its varied applications. Many systems have been proposed for recognition of printed as well as handwritten characters, for Devanagari (Karwankar and Bhalchandra (2010); Desai and Malik(2011); Desai et.al.(2011); Raj et.al.(2013); Holambe et.al.; Aggarwal et.al.(2012); Malaviya et.al.(1996); Dhandra et.al.(2010); Bharath and Madhvanath(2010); Shaw et. al.(2008); Singh et. al.(2011); Chavan et. al.(2013); Koshti and Govilkar; Agrawal et. al.; Gohil et. al. (2012); Holambe et. al.(2010); Rajput and Mishra; Malik and Deshpande (2009); Jangid(2011); Shukla et. al.(2011); Mapari et. al.(2011); Sharma et. al.(2006); Garg et. al.(2011); Pratap and Arya(2012); Sahu et. al.(2012); Murthy and Hanmandlu(2011); Ramana et. al.(2012); Murthy and Hanmandlu(2011); Deshpande et. al. (2007;2008);
Chapter 1: Introduction

Mukherji and Rege(2008;2009); Patil and Ansari(2014); Kakde and Raut(2012); Singh and Tyagi(2011); Ramteke(2010); Ramteke and Melhotra(2008); Jayadevan et. al.(2011); Kapoor et. al.(2002); Kumar et. al.; Rathi et. al.(2012); Khobragade(2013); Bajaj et. al.(2002); Arora et. al.(2008; 2009; 2011); Kamble and Kamble ((2011); Kumar(2009; 2010); Kumar et. al. (2010;2012); Asthana et. al.(2011); Kompalli et.al.(2009); Shelke and Apte(2011); Vaidya and Bombade(2013); Bhattacharya and Chaudhuri(2005); Ladwani and Malik(2010); Agnihotri(2012); Bansal and Sinha(2000); Kumar and Sengar(2010); Dongre and Mankar(2010); Rani and Kumar(2013)), for Bengali (Sarkar and Biswas(2010); Majumdar(2007); Das and Yasmin(2006); Shukla et. al.(2011); Parui et. al.(2008); Bag and Harit(2013); Bhattacharya and Chaudhari(2005)), for English (Talele et. al.(2011); Koerch et. al.(2010); Choudhary et. al. (2010); Dhandra et. al.(2006); Romero et. al.(2007); Patel et. al.(2012); Pradeep et. al.(2010); Hull et. al.(1990); Vaid and Gupta(2002); Prema and Reddy(2002); Biswas and Parekh(2012); Sharma et. al.(2012); Asthana et. al.(2011)), for Marathi (Ajmire and Warkhede (2010); Jane and Pund(2012); Mahender and Kale(2011); Rajput and Mali(2010); Kale et. al.(2014); Tapkir and Shelke(2012); Patil et. al.(2011); Ajmire et. Al.(2012); Jayadevan et. Al.(2011); Mali(2012); Shelke and Apte(2010;2011); Pawar and Gaikwad(2014)), for Gujarati (Desai (2012); Baheti et. al.(2011)), for Gurumukhi (Kumar and Jindal(2012); Singh and Budhiraja(2012); Singh and Dhir(2012); Kumar and Sengar(2010)), for Kannada (Dhandra et. al.(2009;2010;2011); Acharya et. al. (2008); Niranjan et. al.(2009); Sangame et. al.(2009); Vaidya and Bombade(2013)), for Telugu (Dhandra et. Al. (2009;2010); Jawahar et.al. (2003); Rao et.al.(2013); Rajashekarardhya and Ranjan (2008); Asthana et.al.(2011)), for Malayalam (Chacko and Anto(2010); Rajashekarardhya and Ranjan(2008)), for Hindi (Jawahar et.al.(2003); Hanmandlu et.al.(2007); Garg et.al.(2010;2011;2013)), for Arabic (Chun et.al.(2009); Abd(2007)), for Chinese (Liu et.al.(2010)), for Tamil (Aparna et.al.(2004); Gandhi and Iyakutti(2010); Kannan and Prabhakar(2008);
Chapter 1: Introduction

Rajashekararadhya and Ranjan (2009); Asthana et.al. (2011), for Farsi (Reza et.al. (2011)), for Urdu (Asthana et.al. (2011)) and for Oriya (Bhattacharya and Chaudhari (2005)). Also many systems have been proposed for numeral recognition of different script (Holambe et.al.; Ashoka et.al. (2012); Aggarwal et.al. (2012); Dhandra et.al. (2010); Romero et.al. (2007); Das and Yasmin (2006); Rajput and Mali (2010)).

Pal and Chaudhari (2001) presented in their brief survey on Indian script recognition sufficient amount of work is reported for printed and handwritten character recognition. Also reported the status of present research and presented scope for future work which consists of OCR for poor quality documents, multi font OCR, multi script OCR, handwritten OCR and OCR for the visually handicapped.

Aarti Desai et. al. (2011) proposed a system for handwritten Devanagari character recognition. They have used minimum edit distance classifier and combination of chain code, branch point and end point features. Using the combination of these features is reported 87 recognition accuracy for 150 characters.

Chavan S. V. et. al. (2013) presented a system for recognition of handwritten compound Devanagari characters. Moment base feature extraction techniques are used to extract geometric features and Zernike moment features. MLP and k-NN classifiers are used for classification and recognition accuracies of 98.78% and 95.65% using MLP and k-NN classifier respectively are achieved on a database of 27000 basic and compound characters.

Karbhari V. Kale et. al.(2014) presented a Zernike moment based feature extraction technique for handwritten Marathi compound characters. Database of 9600 basic characters, 9000 compound characters and 3000 split characters has been developed. Local structural classification and zone based zernike moment features are extracted. Recognition is carried out by using SVM and k-NN
Chapter 1: Introduction

classifiers, where 98.37% recognition accuracy is achieved by SVM classifier and 95.82% accuracy by k-NN classifier.

Malik and Deshpande (2009) presented a novel approach for printed and handwritten Devanagari characters by using regular expressions in finite state models. Recognition accuracy reported for printed Devanagari characters is 100.

Shelke and Apte (2010, 2011) have suggested novel approach for handwritten Devanagari compound character recognition consisting of multi-feature and multi-classifier scheme. Database of 35000 character samples has been developed. Structural classification, random transform, wavelet transform, density, Euclidean distance, modified wavelet transforms are used as feature extraction techniques. MLP and Neural network are used for classification. Recognition accuracies reported are 94.22% when wavelet transform is used; 96.23% when modified wavelet transform is used, while for a combination of modified wavelet transform, density and Euclidean distance gives 97.95% recognition accuracy.

Bhattacharya and Chaudhari (2005) presented a brief survey on databases for research on recognition of handwritten characters of Indian script. Databases of 22556 samples of Devanagari numerals, 12938 samples of Bangala numerals, 5970 samples of Oriya numerals have been developed. Database of Devanagari numerals is collected from 1049 users. Also 556 users have written Bangala and Oriya numerals.

Sandhya Arora et. al. (2009, 2011) reported multiple feature and multi classification approach for handwritten Devanagari character recognition. Shadow features, view based features, chain code and moments are used as features for recognition. Neural network classifier is used sequentially for classification using multiple features. Recognition accuracy reported is 90.74% when shadow features and chain code features are used.

Naresh Kumar Garg et. al.(2010) presented a segmentation method using vertical and horizontal projection. Databases of 200 lines and 1380 words of
Hindi text were developed and results of 91.50% for line segmentation, 98.10% for word segmentation, 79.12% for consonants segmentation and for modifiers 86% were reported.

Pal and Chaudhari (2001) presented a segmentation method for printed and text line identification and the segmentation accuracy achieved 98.60% accuracy.

Ajay Talele et. al. (2011) reported a system for handwritten legal amounts written in English. Cavity and closed loop features are used for the recognition purpose. They also have reported 92.50% recognition accuracy.

Alessandro L. Koerich et. al (2013) proposed a system for verification of unconstrained handwritten English words at character level. A database of 85092 English handwritten words is used for the experiments and recognition accuracy is improved by 3.9%.

Bikash Shaw et. al. (2008) made significant contributions towards offline handwritten Devanagari word recognition. They have developed a database of 39700 word samples for offline handwritten Devanagari words, consisting of 100 words. Both Holistic and segmentation based approaches are used for recognition purpose. Chain code, 8 scaler, histogram and zone based features are are extracted and HMM classifier is used for classification purpose. Using holistic based approach 80.2% recognition accuracy is reported, while 81.63% recognition accuracy is reported for segmentation based approach.

Brijmohan Singh et. al. (2011) proposed a novel approach for handwritten Devanagari word recognition using curvelet transform. Database of 28500 samples for handwritten Devanagari word from 30 classes and database of 31860 samples for handwritten Devanagari characters from 46 classes were developed. Curvelet transform and character geometry is used to extract features and recognition accuracy is compared using SVM and k-NN classifier. Recognition accuracy for words is 85.60% using SVM classifier and 93.21% using k-NN classifier.
Gang Liu et. al. (2010) reported a novel approach for handwritten Chinese words. Database of 44208 samples of words has been developed. Holistic approach for recognition is used. LDA and MQDF classifiers are used for classification purposes. Recognition accuracy reported for Chinese words is 91.96%.

R. Jayadevan et. al. (2011) presented a database and a recognition approach for handwritten Devanagari legal amount words. A database of 26720 word samples is developed which contains all Devanagari legal amount words. Gradient, structural features and cavity binary vector matching (BVM) is used for recognition and achieved 80.65% recognition accuracy. A second approach using vertical projection and dynamic time wrapping (DTW) is reported with recognition accuracy 76.69%.

Tapkir and Shelke (2012) reported OCR for handwritten Marathi script. Projection methodology is used for line segmentation and word segmentation. Density feature and Euclidean minimum edit distance classifier is used for recognition. Reported result for line segmentation is 100% and for word segmentation 98%. Recognition accuracy achieved is 92.77% for handwritten Marathi script.

Veena Bansal and R.M.K. Sinha (2000) presented a complete Devanagari OCR system and tested it with real-life printed documents of varying size and font. Most of the documents used were photocopies of the originals. Recognition accuracy reported is 90%.

Neha Avhad et. al. (2015) elaborated system for handwritten Devanagari character recognition. The system addresses the segmentation of handwritten Devnagari text document, the most popular script of Indian sub – continent into lines, words and characters. They have used artificial neural network technique to design to pre-process, segment and recognize Devanagari characters.
Priyanka Kulkarni (2015) et. al. presented brief review on Marathi and Sanskrit word recognition using genetic algorithm. They have used dictionary based approach and curvelet transform features are used for recognition purpose.

Kapil Bamne and Neha Sharma (2015) presented a system for offline classifier for handwritten Devanagari script recognition. They have focused on the recognition of offline handwritten Hindi characters that can be used in common applications like commercial forms, bill processing systems, bank cheques, passport readers, offline document recognition generated by the expanding technological society.

Snehal S. Patwardhan and R. R. Deshmukh (2015) reported a brief review on offline handwritten recognition of Devanagari script. They have elaborated detailed overview of different feature extraction and classification techniques for recognition process Devanagari script by the researchers over the past few decades.

From literature it has been observed that, due to non availability benchmark database of handwritten words, experiments are performed on varied number of samples. Very few experiments were performed on large databases. Many researchers are considering holistic approach for word recognition, in which dataset is limited. Analytical approach for word recognition is segmentation based approach. There are many hazards in segmentation based approach which reduces recognition accuracy. Also many characters are similar in shape and presence of compound characters in some scripts complicates the process of word recognition. It may be concluded that, development of handwritten OCR is most challenging and fascinating task for researchers working in pattern recognition.
Chapter 1: Introduction

1.3 Motivation for the present work, Problem statement:

Marathi is a well known language spoken by the people of Maharashtra. It is written in Devanagari script which is third most widely used script in the world. There are around 100 million speakers of Marathi language which is the fourth largest number of native speakers in India.

Handwritten Marathi OCR has numerous applications like the reading machines for blind and visually impaired, number plate recognition, for reading invoices, postal automation, automated processing of bank cheque and bank statements, digitization of 7/12 documents and ration cards, automated evaluation of answer sheets, automated processing of admission forms and recommendation forms.

Significant work has been reported for handwritten Devanagari/Marathi character recognition and for printed Marathi OCR. However, handwritten Marathi word OCR is not addressed satisfactorily in case of unconstrained handwritten Marathi words. OCR for unconstrained handwritten Marathi word is very complex due to many reasons as stated below:

1. Number of vowels and consonants in Marathi is large.
2. Word formation in Marathi is complex.
3. Vowels can be combined with consonants in forms.
4. Diacritic marks can be placed to the left or right or above or bottom of the consonant.
5. Vocabulary is very large.
6. Marathi has fused characters also known as ‘Jodakshare’.
7. Number of ‘Jodakshare’ s are more and are used frequently as compare to other languages written in Devanagari script.
8. Some of the vowels and consonants are very similar in shapes and structure.
Chapter 1: Introduction

9. Every consonant when it combined with consonant takes form of half character.

10. Literature review shared that not much research is reported for handwritten Marathi word recognition.

The goal of optical character recognition is to come up with a recognizer which has best possible recognition accuracy. In this work we are designing such type of recognizer. **Hence the problem may be stated as:** Given a character set and a database of handwritten characters from the character set, design efficient recognizer that recognizes all characters in the character set accurately.

Efficient recognizer is the recognizer which recognizes handwritten characters using minimum number of features.

Accurate recognition can be defined as high recognition accuracy across all handwritings.

For this problem we have chosen:

1. Character set consists of either Marathi character or Marathi words.
2. A Marathi word set is infinite since meaning of the words is not considered in the present work.

**1.4 Organization of thesis:**

This thesis is organized into eight chapters.

In chapter 2, we are presenting objectives of the research work and brief description of the proposed system to recognize handwritten Marathi word.

In chapter 3, we presented a brief description about Marathi; characters used and the formation of Marathi words. The method of development of database for handwritten Marathi words is elaborated. Also development of database for handwritten Marathi simple words, compound words and isolated characters is presented. Preprocessing techniques that were used to improve quality of word images and to reduce noise are elaborated.
Chapter 1: Introduction

In chapter 4, the methodology for segmentation of handwritten Marathi words was described. In this chapter we have described difficulties in segmentation of handwritten Marathi word. Segmentation algorithms are presented for handwritten Marathi simple and compound words and results are compared with earlier work.

In chapter 5, we are presenting a multilevel classification technique which categorized Marathi characters into six different groups depending upon their special properties.

In chapter 6, feature extraction techniques are presented for handwritten Marathi characters such as zone based symmetric density, moment invariant, zernike moment, discrete wavelet transformations, diagonal, horizontal and vertical features and normalized chain code. Finally we have discussed how to create a knowledge base which contains feature vectors for every image and corresponding class labels.

In chapter 7, classification process is described in detail. Methods used for classification such as k-NN and SVM are described. For rigorous testing and validation a fivefold cross validation technique is presented. A comparative study of two the classifiers namely, k-NN and SVM is elaborated.

Finally, the chapter 8 contains summary, conclusions and future directions of work carried out in this thesis. The results of all the methods proposed in this thesis are compared. Further, the comparative study of proposed method and other methods in literature is also carried out. Lastly, future directions for research based on the present work are presented.