Chapter 3
DEVELOPMENT OF A DATABASE OF HANDWRITTEN MARATHI WORDS, ISOLATED CHARACTERS AND PREPROCESSING
Chapter 3

Development of a database of the handwritten Marathi words, isolated characters and pre-processing

3.1 Introduction
3.2 Marathi characters
3.3 Formation of Marathi words
3.4 Database Development
   3.4.1 Database development for handwritten Marathi simple words
   3.4.2 Database development for handwritten Marathi compound words
   3.4.3 Database development for isolated handwritten Marathi characters
3.5 Pre-processing

In this chapter, we are presenting brief description about Marathi language, characters used in Marathi and formation of Marathi words. Also we have presented the method for development of a database of the handwritten Marathi isolated characters, simple words and compound words. Further pre-processing techniques used to improve quality of word images and to reduce noise are elaborated. Normalization is carried out for handwritten Marathi words and isolated characters without disturbing aspect ratio.

Chapter 3: Development of database for handwritten Marathi words, characters and preprocessing

3.1 Introduction:

An overview of Marathi language, characters used in Marathi and various word formations rule complicate the process handwritten Marathi word recognition. The method of preparing the database for isolated handwritten Marathi characters and isolated handwritten Marathi words is discussed in detail. Pre-processing techniques helps to improve the quality and reduce the noise in the images. Sufficient amount of work had been carried out and reported on isolated Devanagari characters in the literature discussed below.

Bikash Shaw et. al. (2008) have reported a database of 39700 samples using 100 classes from 436 writers for handwritten Devanagari words. Brijmohan Singh et. al. (2011) has developed a database of 28500 word samples for 30 classes for handwritten Devanagari word from 950 writers. Laurent Guichard et. al. (2010) have reported a database of 2000 samples for 10 classes of Devanagari numerals written in word form from one to ten numerals. This database consists of 10 classes and for each class 200 samples are stored. Naresh Kumar Garg et. al. (2010, 2011, 2013) developed a database for handwritten Hindi text consisting of 200 lines and 1380 words. R. Jayadevan et.al. (2011) developed a database of 26720 word samples for handwritten Marathi legal amounts consisting of 114 classes. G.G.Rajput et. al. (2010) used 100 blocks of handwritten Hindi script. Rajiv Kumar et. al. [113] developed a database of 2,000 constrained and 2,000 unconstrained handwritten Devanagari words. Sandip N. Kamble et. al.(2011) developed a database of 100 handwritten Devanagari words. Vijaya Rahul Pawar et. al. (2014) developed a database of 3000 handwritten Marathi words.

It is observed from literature that experiments by researchers were performed on databases various sizes ranging from 100 to 39700 having different datasets. The method for database development and pre-processing techniques applied are discussed in the next sections.
Marathi language: Marathi is well known language spoken by people of Maharashtra. Marathi belongs to the Indo-Aryan group of languages. Indo-Aryan languages are originated from Sanskrit. Currently balbodh script is used for Marathi language and is originated from Devanagari script. Marathi has influence of other languages like Sanskrit, Kannada and Telugu. Also lots of words are entered into Marathi from Persian, Turkish and Arabic as well as Portuguese and the British have influenced Marathi through their words.

3.2 Marathi Characters:

Marathi consists of a total 53 characters out of which 16 are vowels and 37 are consonants.

Marathi vowels:

The 16 Marathi vowels are classified into two groups; the first group contains 12 vowels as shown in Fig. 3.1 while the second group contains four vowels as shown in Fig. 3.2.

First group of vowels are commonly used where as second group of vowels are very rarely used. Out of four vowels of the second group two vowels ( ) have never been used in Marathi and remaining two vowels are found only in three words called 'kL^iptee', 'R^ishI', 'R^itU' as shown in Fig. 3.3.
Chapter 3: Development of database for handwritten Marathi words, characters and preprocessing

![Image](https://example.com/image1.png)

**Fig. 3.3:** Three words contains remaining two vowels

Since second group of Marathi vowels are not commonly used, we have considered only first group of vowels for our study.

**Marathi consonants:**

The 37 consonants are broadly classified into six groups as shown in Fig. 3.4 according to their pronunciation.

![Image](https://example.com/image2.png)

**Fig. 3.4 (a):** First group of Marathi consonants  
**Fig. 3.4 (b):** Second group of Marathi consonants  
**Fig. 3.4 (c):** Third group of Marathi consonants  
**Fig. 3.4 (d):** Fourth group of Marathi consonants  
**Fig. 3.4 (e):** Fifth group of Marathi consonants  
**Fig. 3.4 (f):** Sixth group of Marathi consonants

Out of all 37 consonants first 25 consonants are classified into five groups where each group contains five consonants. First group of consonants is called ‘Kantha’ because they are pronounced from the throat. Second group of consonants is called 'Murdhanya' because they are pronounced by touching the tongue to 'Murdhani' which is a part of the upper jaw between the roof and the teeth. Third group of consonants is called 'Taalavya' because they are pronounced by touching the tongue to the palate. Fourth group of consonants is called 'Dantya' because the tongue touches the teeth while pronouncing these. Fifth group of consonants is called 'Aushthya’ because they are pronounced by touching the lips together. Sixth group consists of twelve remaining consonants which are pronounced using combination of usage of tongue.
In Marathi out of 37 consonants, 36 consonants are commonly used but one consonant (क) is rarely used hence only 36 consonants are considered in this work.

Finally the total 48 characters in Marathi are considered in this work which consists of 12 vowels and 36 consonants as shown in Fig. 3.5.

3.3 Formation of Marathi Words:

In Marathi, word formation is a very complex system because number of vowels and consonants are more than English. Also during word formation vowels may be combined with consonants. Whenever vowels are combined with consonants they will take different forms called diacritic marks such as ‘Kana’, 'Matra’, 'Ukar’, 'Velanti’, 'Anuswar’ or 'Visarg’.

In addition to this Marathi has a complex system of compound or fused characters where more than one consonant are combined called ‘Jodakshare’. In ‘Jodakshare’ first consonant converted into half form and second has its full form. Also two different words are also combined when second word is starting with vowel.

Following Fig. 3.6 shows full consonant and its corresponding half consonant.
Fig. 3.6(a): Group one consonants and its corresponding half consonant.
Fig. 3.6(b): Group two consonants and its corresponding half consonant.
Fig. 3.6(c): Group three consonants and its corresponding half consonant.
Fig. 3.6(d): Group four consonants and its corresponding half consonant.

According to their form half consonants are classified into 5 groups. Group one contains those consonants are having vertical bar at the end and we get corresponding half consonant form by removing vertical bar. Group two is consonants not having vertical bar and we get corresponding half consonant form by adding slanting line below the consonant. Group three contains only one consonant with small vertical line on right-top end (ढ़) and we get half consonant form by removing that small vertical line (ढ़). Group four contains two consonants (क़ and फ़) having curve on the right side and formation of half
Chapter 3: Development of database for handwritten Marathi words, characters and preprocessing

Consonant is obtained by removing half curve. Group five contains consonants which take multiple forms depending on the next character (र and च).

3.4 Database development:

In order to develop a system for offline handwritten Marathi word recognition sufficient amount of database is required. Database has to contain large vocabulary and variations. No standard database is available for Marathi.

3.4.1 Database for Handwritten Marathi simple words:

It is observed from literature that experiments by researchers were performed on databases various sizes ranging from 100 to 39700 having different datasets. Also literature review indicates that benchmark database for handwritten Marathi word is not available for carrying out experiments. Since a benchmark database is not available [32] our first attempt was to develop a database for handwritten Marathi words.

Marathi contains two types of words such as simple words and compound words. Simple words do not have ‘Jodakshare’ while compound words have. To develop a database for handwritten Marathi simple words, a dataset consisting of 50 commonly used Marathi words were selected as shown in Fig. 3.7. While selecting simple words we have taken care that all possible combinations of vowel modifiers and consonants will appear in the words.
To develop a database for handwritten Marathi words, three A4 size sheets were specially designed as shown in Fig. 3.8. These sheets were distributed to 50 writers of different age groups and professions which include students, clerks, teachers etc. There are no constraints imposed on writers, except that they have to write words in the given boxes. Every writer has to write a word for 10 times. Finally, a database of 20210 handwritten Marathi simple word samples of 50 classes written by 50 different users was ready to carry out the experiments.
Chapter 3: Development of database for handwritten Marathi words, characters and preprocessing

Fig. 3.8: Sample of sheets for collection of handwritten Marathi simple words

The handwritten data sheets were then scanned using a flat bed scanner at a resolution of 1200 dpi and stored as gray scale images. Handwritten words from the scanned gray scale images were manually cropped and stored in respective class folders. The Fig. 3.9 shows some handwritten simple words in gray scale cropped from the scanned image of a datasheet.

Fig. 3.9: Sample handwritten Marathi simple words
3.4.2 Database for handwritten Marathi compound words:

As discussed in section 2.4.1 Marathi consists of compound words also called as fused words. Compound word contains fused characters known as ‘Jodakshare’. Occurrences of compound characters in Marathi is more frequent (11 to 12%) as compared to other languages written in Devanagari (5 to 6%) Shelke and Apte (2010).

To develop a database for handwritten Marathi compound words, the dataset consisting of 47 commonly used Marathi words are selected as shown in Fig. 3.10. While selecting the compound words all possible combinations of vowels, modifiers and consonant clusters are considered.

![Dataset of handwritten Marathi compound words](image)

**Fig. 3.10: Dataset of handwritten Marathi compound words**

Also A4 size sheets were specially designed to collect handwritten Marathi word from 50 different users as shown in Fig. 3.11.
Chapter 3: Development of database for handwritten Marathi words, characters and preprocessing

Fig. 3.1: Sample A4 sheet for Handwritten Compound words

We have adopted the same procedure for scanning, manually cropping and storing in the respective class-folders as described in section 2.4.1. Finally, a database of 16073 handwritten Marathi compound word samples of 47 classes written by 50 users is ready for experiments. Sample handwritten compound words are shown in Fig. 3.12.

Fig. 3.12: Sample handwritten Marathi compound words
3.4.3 Database for isolated handwritten Marathi characters:

We have designed A4 size sheets for data collection of Marathi handwritten Marathi characters and distributed the sheets amongst 20. We have adopted same procedure for scanning, cropping and storing into respective class folders as described in Section 3.4.1. Finally, a database of 9600 isolated handwritten Marathi characters of 48 classes, written by 20 persons is ready for experiments.

![Sample A4 sheet for isolated handwritten Marathi characters](image)

Fig. 3.13: Sample A4 sheet for isolated handwritten Marathi characters
3.5 Pre-processing:

Pre-processing refers to a number of operations that may be performed on the input intensity images to obtain outputs with good quality intensity images. The main objective of pre-processing is to remove noise from images, to enhance quality of input images and to represent word images in standard plane.

3.5.1 Noise Removal:

Digital images contain noise due to various reasons such as movement in acquisition process or inaccuracy in instrument for digitization. Noise removal is an important step in preprocessing. There are several techniques for noise removal like low-pass, high-pass, band-pass, spatial filtering, mean filtering, median filtering. To reduce the blurring of word edges, suppress noise and improve some features, the median filter is used. Median filter preserves edges and removes noise. An example of the median filtering process on raw input image is shown in Fig. 3.14.

![Fig. 3.14 Example of Median filtering](image-url)
3.5.2 Binarization:

Binarization operation takes input as indexed, intensity or RGB images and outputs binary images. Here gray scale image is converted into binary image having values 0 and 1. Gray scale image values are converted to 0 and 1 depending upon a threshold. The threshold for the gray scale image is calculated by using histogram-shape based image thresholding suggested by Otsu. Otsu’s method reduces interclass variance.

Ostu’s method assumes two classes of pixels (foreground and background) in input image and calculates the optimum threshold value for separating those two classes. The output binary image contains 0 if values of pixels in input image are less than the calculated threshold value and 1 for all other pixels.

Ostu’s Algorithm:

**Input:** Handwritten Marathi word / character images.

**Output:** Pre-processed handwritten Marathi word / character

**Procedure:**

1. Compute the normalized histogram of the input image. Denote the components of the histogram by \( P_i \), \( i=0,1,2,...,L-1 \).

2. Compute the cumulative sums, \( P_1(k) \), for \( k=0,1,2,...,L-1 \), using

\[
P_1(k) = \sum_{i=0}^{k} P_i
\]

3. Compute the cumulative means, \( m(k) \), for \( k=0,1,2,...,L-1 \), using

\[
m(k) = \sum_{i=0}^{k} i * P_i
\]

4. Compute the global intensity mean, \( m_G \), using

\[
m_G = \sum_{i=0}^{L-1} i * P_i
\]
Chapter 3: Development of database for handwritten Marathi words, characters and preprocessing

5. Compute the between-class variance, \( \sigma_B^2(K) \), for \( k = 0, 1, 2, \ldots, L-1 \), using

\[
\sigma_B^2(K) = \frac{[m_g P_1(k) - m(k)]^2}{P_1(k)[1 - P_1(k)]}
\]

6. Obtain the Otsu threshold, \( k^* \), as the value of \( k \) for which \( \sigma_B^2(K) \) is maximum. If the maximum is not unique, obtain \( k^* \) by averaging the values of \( k \) corresponding to the various maxima detected.

7. Obtain the separability measure, \( \eta^* \), by evaluating

\[
\eta(K) = \frac{\sigma_B^2(k^*)}{\sigma_E^2}
\]

3.5.3 Normalization:

Handwritten words are varying in size and shape. We need to map these word images onto a standard plane (with predefined size) so as to give a representation of fixed dimensionality for classification. Normalization is performed on the image to reduce the inter-class and intra-class variations of the shapes of the words. Normalization operation facilitates segmentation process and improves their segmentation accuracy. Linear normalization method is used to standardize the word images. The standard plane is considered as a square of size 60 pixels x 90 pixels. The width and height ratio of the word image is not disturbed due to normalization.

3.5.4 Thinning:

A morphological operation known as thinning, is also performed on word images. The goal of character thinning is to remove pixels so that an object without holes shrinks to a minimally connected stroke, and an object with holes shrinks to a ring halfway between the hole and outer boundary. Thinning Marathi words is very difficult task due to the presence of loops. This thinning operation preserves Euler number. Thinning operation is related to hit-or-miss transform
Chapter 3: Development of database for handwritten Marathi words, characters and preprocessing

and is represented as follows. Thinning of a set $A$ by a structuring element $B$, is defined as follows in terms of hit-or-miss transforms

$$A \otimes B = A - (A \circ B) = A \cap (A \circ B)^C$$

In this process we have used a sequence of structuring elements as follows:

$$\{B\} = \{B^1, B^2, B^3, ..., B^n\}$$

In the next chapter algorithms for segmentation of words into isolated characters are described and analyzed.