CHAPTER 3

DATA COLLECTION, DIGITIZATION AND PRE-PROCESSING

There are three chief prefatory steps such as data collection, digitization, and pre-processing to proceed in recognition of offline handwritten words. The following sections describe the tasks done to collect data of offline handwritten Gurumukhi words, their digitization, and pre-processing operations. Section 3.1 describes the process of data collection; section 3.2 outlines the digitization phase, and section 3.3 focuses on the pre-processing phase of the present offline handwritten word recognition system. Section 3.4 provides the summary of the complete chapter.

3.1 DATA COLLECTION

A benchmark database plays a significant role in order to perform experiments as well as to make comparisons among the proposed approach and the existing approaches. Due to the non-availability of a public benchmark database in Gurumukhi script, a database was created to perform experiments to recognize offline handwritten words. For the creation of the database, handwritten words were collected from 100 distinct writers in offline mode. This database involves 100 distinct word classes that correspond to place names of Punjab state in Gurumukhi script. Each of the 100 different writers wrote each Gurumukhi word 10 times that resulted in a total of 1,00,000 samples of handwritten words (100×100×10=1,00,000) as depicted in Table 3.1. These handwritten words were collected from diverse schools, colleges, and other public places of Punjab state. To generate variations, the considered writers belonged to different age groups who used either black or blue pen for writing the words. The database was collected using A4 sheets as illustrated in Figure 3.1.

Table 3.1. Database description

Script	Number of classes	Number of writers	Number of words written by each writer	Total specimens
Gurumukhi	100	100	10	1,00,000

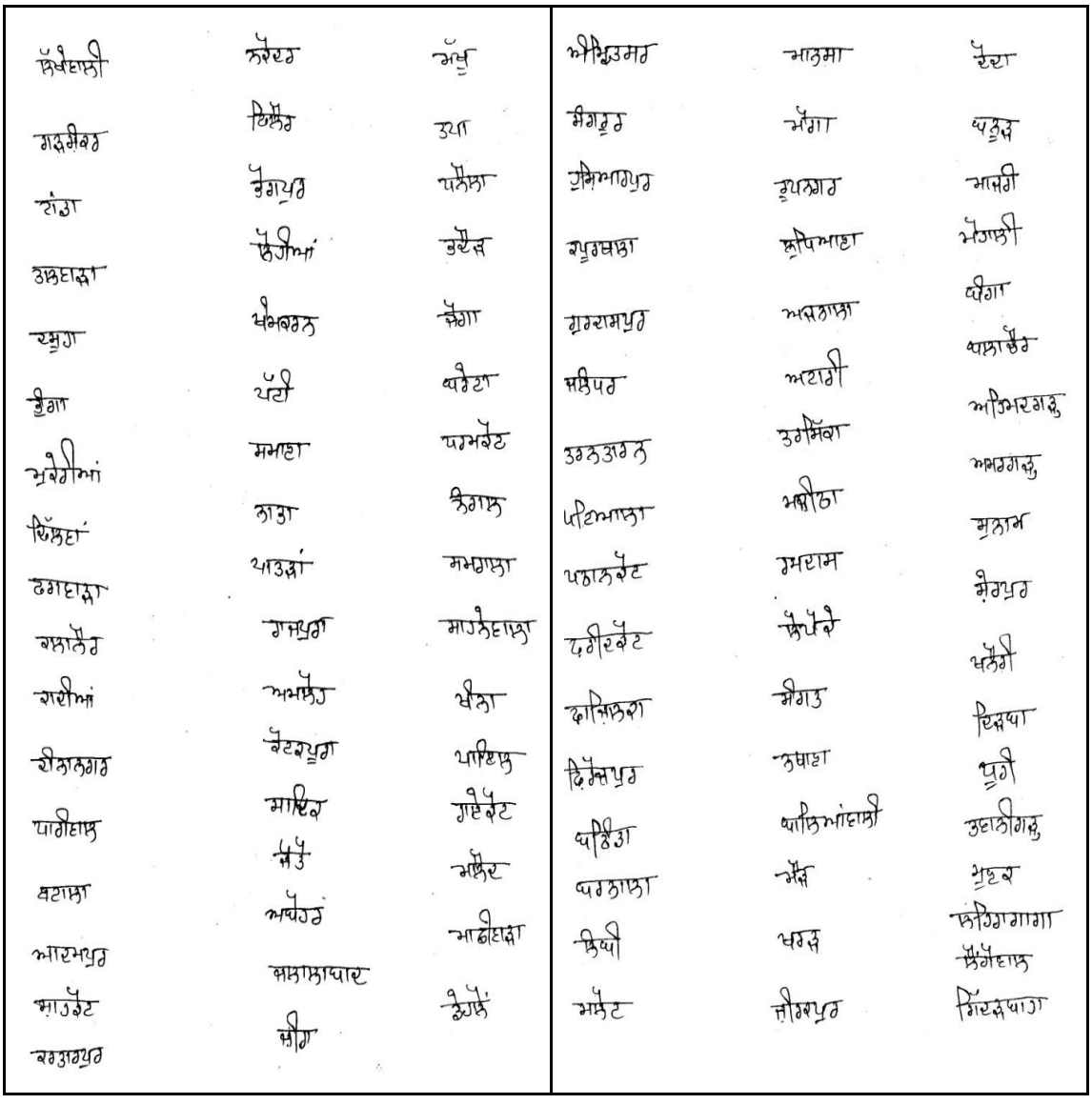


Figure 3.1. Offline handwritten Gurumukhi words

The dataset comprising 40,000 words has already been made available publicly at the link: https://sites.google.com/view/gurmukhi-benchmark/home/word-level-gurmukhi-dataset (Kaur and Kumar, 2019) for the researchers so that they can

evaluate their proposed techniques using this public dataset without generating their own dataset. This dataset has been named HWR-Gurmukhi_Postal_1.0 that includes 100 distinct classes of words (place names). This dataset was created with the help of 40 different writers who wrote each word 10 times, thus resulting in a total of 40,000 words (100×40×10=40,000). This dataset has been partitioned as 70% training and 30% testing set as depicted in Table 3.2.

Table 3.2. Benchmark dataset description

Dataset Name	**Number of classes**	**Number of writers**	**Number of words written by each writer**	**Number of training specimens**	**Number of testing specimens**	**Total specimens**
HWR-Gurmukhi_Postal_1.0	100	40	10	28,000	12,000	40,000

3.2 DIGITIZATION

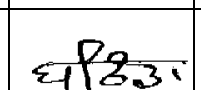
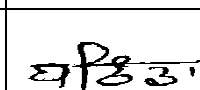
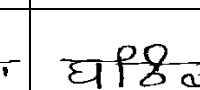
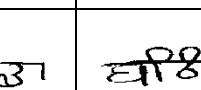
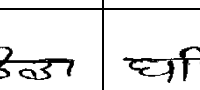
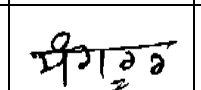
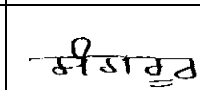
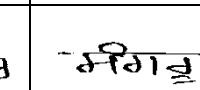
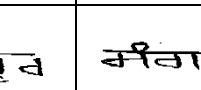
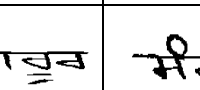
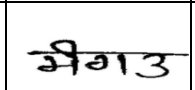
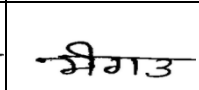
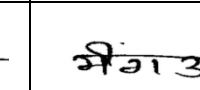
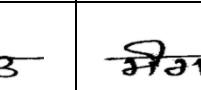
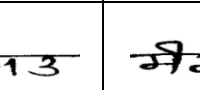
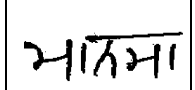
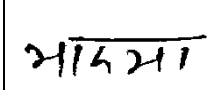
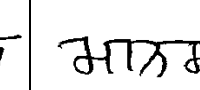
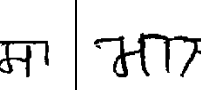
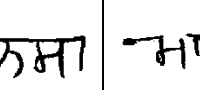
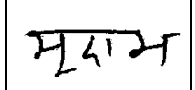
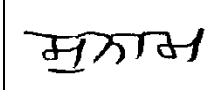
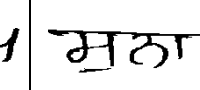
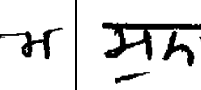
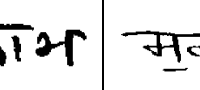
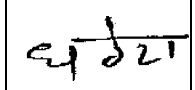
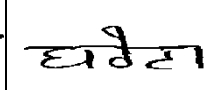
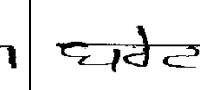
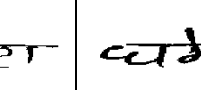
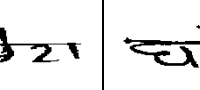
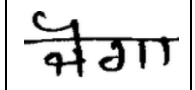
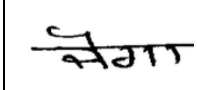
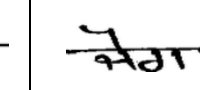
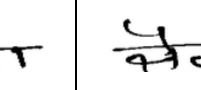
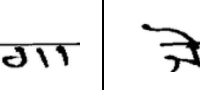
The digitization process is applied to transform paper based documents into electronic forms (digital forms). The digital image of the document is produced by scanning the document through the scanner. All the documents in the present research comprising handwritten words were scanned at 300 dpi (dots per inch) resolution. The scanned documents were stored in a .jpeg image format. After digitization, the produced digital image of the document was given as an input to the pre-processing stage of offline handwritten Gurumukhi word recognition.

3.3 PRE-PROCESSING

In the pre-processing stage, there are various operations that can be performed on the digitized image of the document. Here, in the pre-processing stage, three operations such as binarization, normalization, and thinning operations were carried out. In binarization operation, the threshold constant was placed between higher and lower

values corresponding to white and black, respectively, in order to generate a binary form of the document image. After getting the binarized image, the words were sliced from the document image which was then further cropped to eliminate the additional white space enclosing the word. Words were then normalized into a window of size 256×64 in order to provide uniformity to the different sized words. Table 3.3 lists the few samples of offline handwritten Gurumukhi words written by 5 distinct writers such as W1, W2, W3, W4, and W5. The normalized words were placed in the .bmp image format. After normalization, a thinning operation was applied that decreased the text width from multiple pixels to a single-pixel as shown in Table. 3.4. The most extensively used algorithm for thinning operation, that is a parallel thinning algorithm (Zhang and Suen, 1984), was applied for the present work.

Table 3.3. Offline handwritten Gurumukhi words

Gurumukhi Word	**W1**	**W2**	**W3**	**W4**	**W5**
ਬਠਿੰਡਾ					
ਸੰਗਰੂਰ					
ਸੰਗਤ					
ਮਾਨਸਾ					
ਸੁਨਾਮ					
ਬਰੇਟਾ					
ਜੋਗਾ					

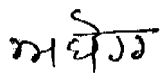
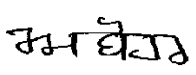
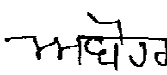
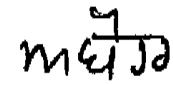
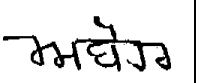
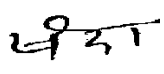
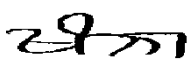
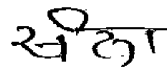
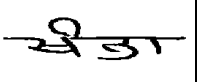
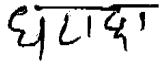
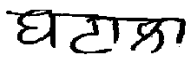
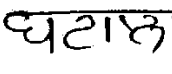
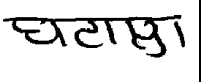
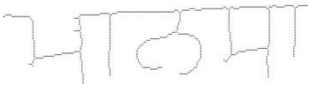
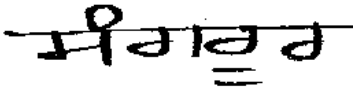
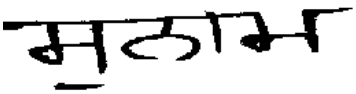
Gurumukhi Word	**W1**	**W2**	**W3**	**W4**	**W5**
ਅਬੋਹਰ					
ਖੰਨਾ					
ਬਟਾਲਾ					

Table 3.4. Thinned images corresponding to digitized images of handwritten Gurumukhi words

Digitized Image	**Thinned Image**
	
	
	
	
	

The whole database has been stored in 100 folders where each folder follows the format i.e. Sr.No._PlaceName where Sr.No. indicates 1 to 100 numbers (that refers to 100 distinct classes) and PlaceName refers to the name of the place whose image has been stored in that folder (For example 1_Bathinda, 2_Lambi, 3_Longowal, 4_Gurdaspur, 5_Zirakpur, etc.). Thus, the whole database has been arranged systematically in order to avoid any kind of confusion.

3.4 CHAPTER SUMMARY

This chapter has discussed the data collection, digitization, and pre-processing stages of the offline handwritten Gurumukhi word recognition system. A database has been developed that comprises 1,00,000 handwritten words in Gurumukhi script which correspond to 100 place names of Punjab State of India. The developed database has been prepared after going through the data collection, digitization, and pre-processing stages. A dataset of 40,000 Gurumukhi words has already been made available by us for public use so that the researchers in the field of word recognition can utilize this benchmark dataset to test their proposed techniques without creating a new dataset from scratch.