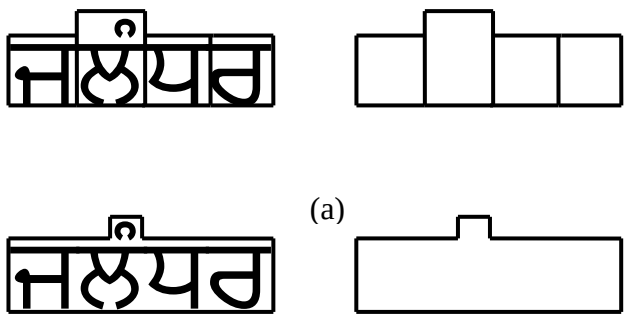
CHAPTER 1

INTRODUCTION

In order to retain the documents for a longer time, it is required to store these documents in digital form. With digitization, historical documents, books, and other such types of documents can be stored in original form for the forthcoming generations. After storage, the documents can be retrieved using pattern recognition algorithms which play an important role in machine vision applications. Optical Character Recognition (OCR) is the area that comes under the pattern recognition field, employed to analyze the documents. OCR is the technique by which the textual symbols on a paper get converted to a format that is processed by a machine for the purpose of recognition. Thus, OCR plays a significant role in the conversion of paper-based documents to the paperless digital (electronic) form. A sequence of characters forms a word and the word recognition approach is the way to recognize the sequence of characters in order to recognize the whole word. A lot of work has been considered in the literature for character recognition and word recognition in many Indic and Non-Indic scripts. But till now no recognized work is available in offline handwritten Gurumukhi word recognition. Moreover, handwritten word recognition finds its application in postal automation. The goal of postal automation is to interpret the handwritten addresses written on the posting envelopes. A lot of work has been considered in the literature for Non-Indic scripts that finds its role in postal automation (Kimura *et al.,* 1995; Srihari and Keubert, 1997; Bartnik *et al.*, 1998; Kim and Govindaraju, 1998; Mahadevan and Srihari, 1999; Wang and Tsutsumida, 1999; Plamondon and Srihari, 2000; Liu *et al.*, 2002; Gao and Jin, 2012; Liu *et al.*, 2014). But there are only a few studies available for Indic scripts (Pal *et al.*, 2006; Wen *et al.*, 2007; Roy, 2008; Pal *et al.*, 2009; Pal *et al.*, 2012, Thadchanamoorthy *et al.*, 2013; Sharma *et al.*, 2017; Roy *et al.,* 2020) and till now, no postal automation system is available for Gurumukhi script. In this thesis, the work is considered to recognize handwritten words (place names) in Gurumukhi script and thus, can be employed to postal automation in Gurumukhi script. Gurumukhi script is utilized to write the Punjabi language which is the official language of Punjab state of India.

1.1 BACKGROUND OF WORD RECOGNITION SYSTEM

Word recognition is the approach to recognize words, which may be printed or hand-written using natural handwriting. There is the existence of many models to recognize words. According to the oldest model in the literature, words are considered as a whole unit rather than a collection of individual characters. Some view that words are recognized by considering information related to the pattern of ascending, descending, and neutral characters. Cattell (1886) was the first psychologist who proposed the word shape model as the envelope generated by the word outline as exhibited in Figure 1.1 (a, b). He found the "Word Superiority Effect" and observed that the subjects were more accurate to recognize the words as compared to the letters. This finding was first reported by Woodworth (1938) in his influential textbook "Experimental Psychology", which was confirmed by Smith (1969) and Fisher (1975). Gough (1972) proposed a model to recognize a word based on letters that are read sequentially from left to right. This proposed model is compatible with Sperling's (1963) finding that the letters can be recognized at a rate of 10-20 ms per letter. Bouwhuis and Bouma (1979) presented an approach to recognize words on the basis of the probability of recognizing each of the characters that constitute a word. They revealed that the shape of the word can be delineated with respect to the characters in their positions.



(a)

(b)

Figure 1.1. (a) Word shape recognition (b) Word shape recognition using envelope around the word

Based on the mode of data acquisition, word recognition systems can be segregated into two categories, namely, Printed Word Recognition and Handwritten Word Recognition, as exhibited in Figure 1.2.

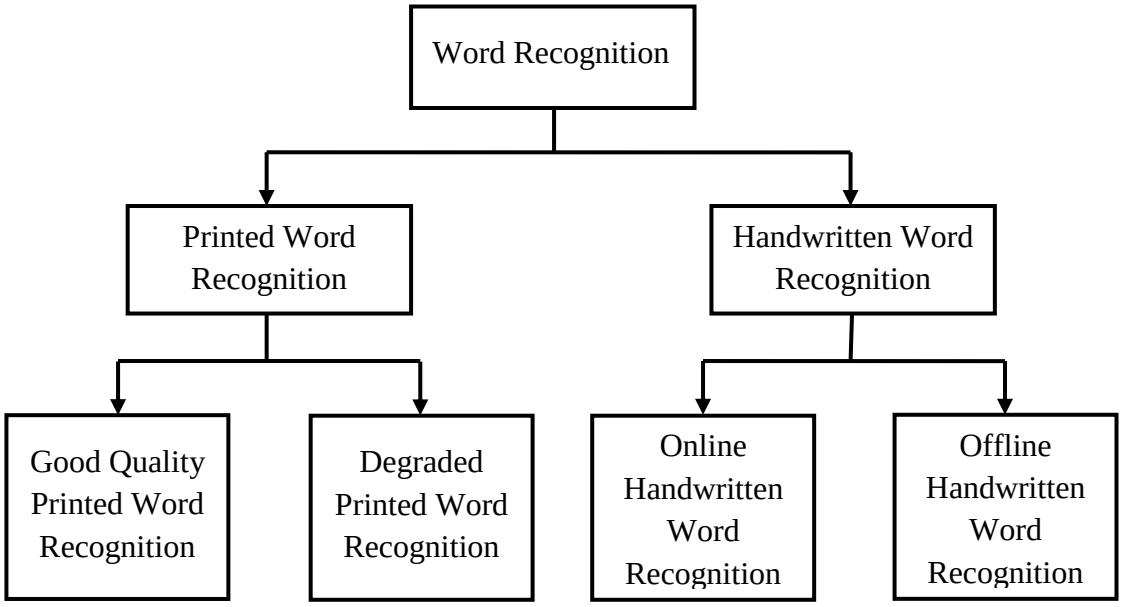


Figure 1.2. Word recognition classification

1.1.1 Printed Word Recognition

Printed Word Recognition considers the machine printed form of documents in order to recognize the words. In this system, the scanner scans the machine printed document to transform it into the digital form. This digital image is then pre-processed to extract the words from it. The required features are extracted from the word images which are then fed to the recognition algorithms for the recognition purpose. There are numerous commercial systems available to recognize printed text. Printed Word Recognition can further be bifurcated into categories, namely, Good quality Printed Word Recognition and Degraded Printed Word Recognition.

1.1.1.1 Good Quality Printed Word Recognition

Good quality printed words are well-printed words that are noiseless in nature. The quality of the input image determines the accuracy of OCR applications (Idris and Panchanathan, 1997; Cattoni *et al.*, 1998; Li *et al.*, 2000).The good quality printed words provide good accuracy and efficiency for the recognition algorithms.

1.1.1.2 Degraded Printed Word Recognition

Degraded printed words are those words that comprise touching, overlapped characters and may also comprise background noise. Various factors lead to these

degradations in a scanned text image, like the flaws in the printing of the text on the sheet, the flaws in feeding the printed text sheet into the scanner, and the flaws in the digitization phase. In order to recognize these degraded printed words, the mentioned issues need to be resolved.

1.1.2 Handwritten Word Recognition (HWR)

Handwritten Word Recognition (HWR) is the approach to recognize words that are written using any natural language. The words are recognized using a machine (computer) after getting a machine-readable format of the documents. HWR mainly necessitates an OCR (Optical Character Recognition) system for the recognition of printed/handwritten text. HWR is divided into two streams, namely, online handwritten word recognition and offline handwritten word recognition as discussed in the following sub-sections.

1.1.2.1 Online Handwritten Word Recognition

In online handwritten word recognition, the words are written on a digitizing tablet using a special pen/pencil which transforms the handwritten words into digital form. The online handwritten word recognition system takes into account the track of pen-tip movements to analyze and recognize a word. Thus, the temporal information, such as position and velocity of the pen inclusive of its track, is available to the recognition algorithms. Most of the algorithms strive to recognize the text as it is being written, so it is also termed as "real time" handwritten word recognition.

1.1.2.2 Offline Handwritten Word Recognition

In offline handwritten word recognition, words are written on a paper sheet using pen/pencil. Then the paper sheet is fed into the computer using a scanner that scans the document to convert it into a digitized image. After getting a digitized image of the document, the recognition algorithms are employed to recognize the handwritten words.

Table 1.1 describes the comparison between online and offline handwritten word recognition and illustrates the superiority of online handwritten word recognition as compared to offline handwritten word recognition in terms of accuracy

and speed of recognition. Thus, offline handwritten word recognition is undoubtedly a challenging task.

Table 1.1. Comparison between Online and Offline Handwritten Word Recognition

Sr. No.	Basis of Comparison	Online Handwritten Word Recognition	Offline Handwritten Word Recognition
1.	Process	It is real time process which seizes the dynamic information of writing.	It is an offline process which captures the static information about the word after it is written on a paper.
2.	Pre-processing	It requires less pre-processing operations.	It requires more pre-processing operations.
3.	Special equipment	The writer requires special equipment such as digitizer to write with special pen/pencil.	The writer only requires paper and pen/pencil without any special equipment.
4.	Suitability	It cannot be applied to already printed/handwritten documents.	It is suitable for already printed/handwritten documents.
5.	Speed	It has higher recognition speed.	It has relatively lower recognition speed.
6.	Accuracy	It supports higher recognition accuracy.	It supports relatively lower recognition accuracy.

1.1.3 Approaches to Handwritten Word Recognition

Document image analysis and recognition is one of the remarkable efforts to make a paperless society. Handwritten word recognition is an active area of research in the document image analysis and recognition field. In order to recognize handwritten

words, there are mainly two approaches available, namely, analytical approach and holistic approach which are discussed in the following segments.

1.1.3.1 Analytical Approach

The analytical approach considers that a word is a collection of individual characters (Bozinovic and Srihari, 1989; Edelman *et al.*, 1990; Blumenstein and Verma, 1999; Roy *et al.*, 2005a; Lee and Verma, 2011; Bouaziz *et al.*, 2014; Jayech *et al.*, 2016; Pramanik and Bag, 2020). Thus, to recognize a word, at first, it recognizes the individual characters by partitioning the word into its primitive characters. As this approach takes into account the segmentation process, it is also known as the segmentation-based approach.

1.1.3.2 Holistic Approach

The holistic approach considers the whole word as an individual entity (Dehghan *et al.*, 2001; Madhvanath and Govindaraju, 2001; Namane *et al.*, 2005; Roy *et al.*, 2005b; De Oliveira *et al.*, 2009; Pal *et al.*, 2009; Acharyya *et al.*, 2013; Patel *et al.*, 2015b; Dasgupta *et al.*, 2016; Bhowmik *et al.*, 2019; Ghosh *et al.*, 2019). Thus, the complete word is recognized as a whole without partitioning the word into its constituent characters. As this approach does not consider the segmentation process, it is also known as the segmentation-free approach.

Sometimes, the characters of a word may touch each other or/and may overlap which leads to an issue in word segmentation. This segmentation issue can be resolved using holistic approach to word recognition.

Handwritten word recognition is a prominent discipline in the domain of document analysis and recognition. In this field, the researchers have been putting their maximum efforts for the last 10 years (Bianne-Bernard *et al.*, 2011; Pal *et al.*, 2012; Acharyya *et al.*, 2013; Bhowmik *et al.*, 2014a; Bhowmik *et al.*, 2014b; Shaw *et al.*, 2015; Dasgupta *et al.*, 2016; Imani *et al.*, 2016; Assayony and Mahmoud, 2017; Tavoli *et al.*, 2018; Arani *et al.*, 2019; Bhowmik *et al.*, 2019; Ghosh *et al.*, 2019; Malakar *et al.*, 2020; Pramanik and Bag, 2020). As per the national readership survey 2012, it has been revealed that in India people read less than 10% of the newspapers in English, whereas 33% of the newspapers are read in Hindi and the rest of the newspapers are read in regional languages. India comprises 22 regional languages and

13 distinct scripts (Obaidullah *et al.*, 2013). So, there is a dire need for offline handwritten recognition systems in regional scripts such as Gurumukhi, Telugu, Bangla, Tamil, Devanagari, etc. to provide services to the people utilizing these regional languages. The work presented in the thesis is an effort in this direction to provide a recognition engine to recognize offline handwritten words in Gurumukhi script using holistic approach. Moreover, handwritten documents can be found at several places like schools, colleges, banks, post offices, etc., which comprise a large amount of handwritten data in the form of signatures, faxes, postal addresses, etc. In this scenario, the research undertaken is considered significant to recognize offline handwritten documents.

1.1.4 Pros of Word Recognition

- Handwritten documents can be saved electronically in the digitized form.
- After getting the digitized form, handwritten documents can be stored for forthcoming generations.
- The handwritten documents can be recognized by machine (computer) without re-entering the data manually into the computer.
- It reduces the human errors associated with manual processing of the documents and boosts up the processing speed.
- It minimizes the number of character errors of the OCR system by using a holistic approach to word recognition.
- The word recognition system finds its applications in various areas like forms automation, signature verification, postal automation, bank cheque verification, writer identification, etc.

1.1.5 Cons of Word Recognition

- Due to the diverse writing styles of individuals, it is laborious to recognize handwritten words.
- The inferior quality of text images and character spaces in a word leads to an issue in word recognition.
- Finding the right segmentation point in words due to touched or/and overlapped characters creates an overhead in recognition as incorrect segmentation results in incorrect recognition.

- For techniques based on prototype matching, a new prototype needs to be generated with each additional word.
- Sometimes the combination of certain adjacent characters creates confusion to the handwritten word identification. In Gurumukhi script, this issue is particularly associated with words that are having characters like **ਖ,ਥ,ਰ,ਗ,** etc. For Example, **ਰਾਮ** and **ਗਮ**

1.2 PHASES OF AN OFFLINE HANDWRITTEN WORD RECOGNITION SYSTEM

The offline handwritten word recognition system generally includes seven phases, namely, digitization, pre-processing, segmentation, feature extraction, feature selection, classification, and post-processing which are illustrated in Figure 1.3.

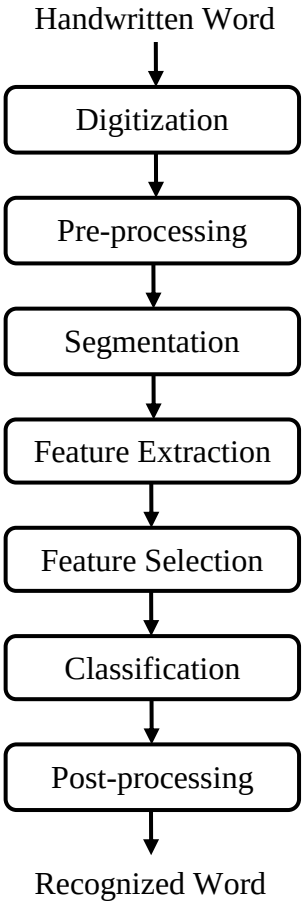


Figure 1.3. Schematic diagram of offline handwritten word recognition system

1.2.1 Digitization

To process offline handwritten documents, these must be saved in machine readable format which is done using the digitization process. In the digitization process, the handwritten documents are fed into the machine using a scanner which scans the documents to get digital images of the documents. This process of converting the paper based documents into digital form (electronic form) is called digitization. After digitization, the bitmap image of the documents is produced upon which the subsequent operations can be performed.

1.2.2 Pre-processing

The documents are written by various writers which may differ due to diverse writing styles. So, to minimize these variations and to eliminate irrelevant information from the documents, pre-processing operations are employed. This phase includes a number of operations on the digitized images of the documents that are obtained through the digitization phase. It mainly comprises three basic operations, namely, binarization, normalization, and thinning operations. Binarization operation is employed to get the binary image of a document in black and white form. To get the binary image, the threshold constant is positioned between lower and higher values that are equivalent to black and white, respectively. Normalization operation is employed to provide uniformity to the words written by different writers. It allocates the same size to the handwritten words. The thinning operation is applied to decrease the text width from multiple pixels to a single pixel. The most widely used algorithm for thinning operation is the parallel thinning algorithm suggested by Zhang and Suen (1984).

1.2.3 Segmentation

The process of partitioning a document into its sub-components is referred to as segmentation. The document can be partitioned into lines, words, and characters. The key point in segmentation lies in finding the right point of segmentation for lines, words, and characters. In word segmentation, the word is partitioned into characters so that the word can be recognized by recognizing its individual characters. Thus, incorrect segmentation can result in incorrect recognition. Segmentation of handwritten words is undoubtedly a challenging task due to the diverse writing styles

of individuals. Thus, for the work carried out by us, the holistic approach (segmentation-free approach) to offline handwritten word recognition has been adopted.

1.2.4 Feature Extraction

The process of extracting the significant characteristics (features) from the word images is called feature extraction. This phase takes into account the relevant shape of the word. This is the major phase for the recognition purpose because based on the extracted features, the words are discriminated from one other. Thus, this phase aims at extricating the most discriminant characteristics from the word images for increasing the rate of recognition. There are mainly two feature extraction techniques available, namely, structural feature extraction technique and statistical feature extraction technique. The structural feature extraction technique takes into account the structural features of the word. These features are extracted on the basis of geometric and topological characteristics of the word by considering local as well as global properties, for example, number of endpoints, number of vertical or horizontal lines, number of cross points, etc. In the statistical feature extraction technique, the statistical characteristics are acquired from the statistical dispersion of pixels (Hallale *et al.*, 2013). Some of the statistical feature extraction techniques are zoning, diagonal, directional, intersection & open-end points features, etc.

1.2.5 Feature Selection

When all the extracted features are utilized for the recognition purpose, then it can lead to the poor performance of the model. Because the presence of extravagant features in the feature set increases the training time of the model and also results in higher storage requirements. So, it becomes essential to select only important features in order to lower the training time as well as to enhance system performance. After the feature extraction phase, the role of the feature selection phase comes into play. Feature selection is a process to select only significant and pertinent features by removing the extravagant features from the original feature set. Due to the diverse writing styles of individuals, feature selection is undoubtedly a challenging task in handwritten documents. This is because it is required to select those features that obtain all the significant features from the handwritten characters/words which aid in

the recognition task. In the literature, there are numerous features selection techniques available such as Consistency Based Analysis (CBA), Correlation Feature Set (CFS), Chi-Squared Attribute (CSA), Principal Component Analysis (PCA), etc.

1.2.6 Classification

Classification phase is considered as the decision making phase because in this phase, decision related to the class of the word image is made. The decision is made according to the extracted/selected features of the word image. Thus, the extracted/selected features from the word image play a significant role to determine the class membership of the word image. The main objective of classification is to develop restraint that can aid in minimizing the misclassification pertinent to feature extraction/feature selection. There are a number of classifiers such as Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Multilayer Perceptron (MLP), Hidden Markov Model (HMM), Decision Tree, Random Forest, Bayesian Network etc. available in the literature for the classification purpose. The effectiveness of any classifier relies on its ability of relating features to the class of the word image.

1.2.7 Post-processing

Post-processing is the last phase in the word recognition system that is used to enhance the overall classification accuracy and to reduce the misclassification rate. Thus, to refine the errors introduced during the classification phase, post-processing is applied. To correct errors, the two most widely used post-processing approaches, namely, dictionary lookup and statistical approach (Lehal and Singh, 2002) can be considered.

1.3 APPLICATIONS OF AN OFFLINE HANDWRITTEN WORD RECOGNITION SYSTEM

- **Form automation:** Forms are part of routine work in various offices which contain useful information in handwritten form. These forms can be automated by storing them in digitized form and can be processed using an offline handwritten word recognition system.

- **Postal automation:** In the era of digitization, postal cards are still used in various government and private offices for sharing some legal information. Postal cards contain handwritten addresses which can be recognized by an offline handwritten word recognition system. The recognition system can be used to read the handwritten addresses and postal codes, thus helpful to sort mails automatically.
- **Writer identification:** Persons differ from one other due to their distinct writing styles. Thus, the distinct writers can be identified based on their writing using an offline handwritten word recognition system. The writer identification is a type of biometric recognition that is beneficial for the verification and authenticity of individuals in various fields like signature identification and verification in banks.
- **Analysis of historical documents:** There are many historical handwritten documents available in the library which need immediate processing. These documents can be processed by offline handwritten word recognition system just by scanning them through the scanner without re-entering the whole document in the system.
- **Bank cheque processing:** The cheques in the bank contain crucial information like account number, account holder name, amount, signature, etc. in a handwritten form which needs to be processed for faster processing. Offline handwritten word recognition system can be employed to verify the account holder signatures and to process all the relevant handwritten information contained in the cheque.

1.4 OVERVIEW OF GURUMUKHI SCRIPT

The word "Gurumukhi" has been derived from the word "Guramukhi" which has the meaning "from the mouth of Guru" which means the words uttered by Guru. The Gurumukhi script is utilized to write the Punjabi language which is the official language of Punjab (Indian state). This script comprises thirty-five consonants, six additional consonants that are created by positioning a dot at the foot of the consonant, nine vowel diacritics, three sub-script letters, and three auxiliary signs as depicted in Tables 1.2-1.5.

Table 1.2. Gurumukhi characters with pronunciation

ੳ Ura	ਅ Era	ੲ Iri	ਸ Sussa	ਹ Haha
ਕ Kukka	ਖ Khukha	ਗ Gugga	ਘ Ghugga	ਙ Ungga
ਚ Chucha	ਛ Chhuchha	ਜ Jujja	ਝ Jhujja	ਞ Yanza
ਟ Tainka	ਠ Thuttha	ਡ Dudda	ਢ Dhudda	ਣ Nahnha
ਤ Tutta	ਥ Thutha	ਦ Duda	ਧ Dhuda	ਨ Nunna
ਪ Puppa	ਫ Phupha	ਬ Bubba	ਭ Bhubba	ਮ Mumma
ਯ Yaiyya	ਰ Rara	ਲ Lulla	ਵ Vava	ੜ Rahrha
ਸ਼ Shusha pair bindi	ਖ਼ Khukha pair bindi	ਗ਼ Gugga pair bindi	ਜ਼ Zuzza pair bindi	ਫ਼ Fuffa pair bindi
ਲ਼ Lulla pair bindi				

Table 1.3. Gurumukhi Vowels

◌ਾ Kanna	◌ਿ Sihari	◌ੀ Bihari	◌ੇ Lavan	◌ੈ Dulavan
◌ੁ Onkar	◌ੂ Dulankar	◌ੋ Hora	◌ੌ Kanaura	

Table 1.4. Sub-script letters

◌੍ਹ pairī̃ hāahāa	◌੍ਰ pairī̃ rāarāa	◌੍ਵ pairī̃ vāavāa

Table 1.5. Auxiliary signs

◌ੰ Tippi	◌ਂ Bindi	◌ੱ Addhak

1.4.1 Properties of Gurumukhi Script

- Gurumukhi script is the 10th most widely used script in the world [Source: Growth of Scheduled Languages: 1971, 1981, 1991, 2001 and 2011, Census of India, Ministry of Home Affairs, Government of India].
- It doesn't have any case sensitivity.
- It follows the left to the right style of writing i.e. it is written in a horizontal way.
- The characters of Gurumukhi script form a word by connecting to one other through the horizontal line at the top of the characters which is called the headline.
- A Gurumukhi word can be partitioned into three zones, namely, upper zone, middle zone and lower zone in a horizontal manner, as depicted in Figure 1.4.

Figure 1.4. Gurumukhi word partitioned into three zones

- The area above the headline depicts the upper zone that includes the vowels.
- The area below the headline and above the baseline shows the middle zone. All the consonants and some part of vowels appear in the middle zone.
- The lower zone is represented below the baseline and includes few vowels and sub-script characters at the foot of the consonants.

1.5 OBJECTIVES OF THE WORK

The objectives of the present work are defined as below.

1. Generation of corpus of Punjab state place names in offline handwritten Gurumukhi script.
2. Investigating and exploring the existing relevant work for analyzing feature extraction and classifiers used for word recognition.
3. To propose and develop new features for classification of offline handwritten Punjab state place names.
4. To propose efficient classifier for recognition of offline handwritten Punjab state place names.
5. To analyze and compare the performance of proposed technique/s with respect to existing methods.

1.6 ASSUMPTIONS

The following assumptions have been taken into account for performing experiments in the present work.

1. The present work only considers pre-segmented handwritten words of Gurumukhi script and doesn't comprise any non-textual items such as figures, images etc.
2. The handwritten words considered for the present work are exempted from noise. There is no need of skew detection and correction.
3. The handwritten documents are scanned at 300 dpi (dots per inch) resolution.

1.7 MAJOR CONTRIBUTIONS AND ACHIEVEMENTS

1. The literature related to word recognition in various Indic and non-Indic scripts has been surveyed.
2. A corpus of 1,00,000 samples of handwritten words (place names) in Gurumukhi script has been generated.
3. A few existing techniques for feature extraction in handwriting recognition have been implemented for handwritten word (place name) recognition.
4. Efficient combinations of different feature extraction techniques have been employed for the recognition of handwritten words (place names).
5. Different feature selection techniques, namely, Consistency Based Analysis (CBA), Correlation Feature Set (CFS), Chi-Squared Attribute (CSA) and Principal Component Analysis (PCA) have been employed to select the meaningful features and to improve the performance of the experimental framework's interpretability.
6. In the classification phase, a hybrid of classifiers has been utilized to improve system performance.

1.8 ORGANIZATION OF THE THESIS

The chief objective of this work was to develop an offline handwritten Gurumukhi word recognition system for postal automation. In order to achieve this objective, various experiments have been performed based on different recognition algorithms. The thesis has been organized as mentioned below.

In this chapter, various phases of offline handwritten word recognition system along with its application areas have been elaborated. An overview of the Gurumukhi

script has also been provided in this chapter. Chapter 2 provides the literature review in various non-Indic and Indic scripts for character and word recognition. The algorithms which have been used in various phases of the experimented recognition system have also been mentioned in detail. Based on the literature review, research gaps have been provided. In Chapter 3, three phases, namely, data collection, digitization, and pre-processing have been considered to proceed for offline handwritten word recognition system. In Chapter 4, an offline handwritten Gurumukhi word recognition system based on holistic approach and Adaptive Boosting (AdaBoost) methodology has been presented. For this work, three features, namely, zoning features, diagonal features, intersection & open-end points features along with their combinations and three classification techniques, namely, k-Nearest Neighbour (k-NN), Support Vector Machine (SVM), and Random Forest have been taken into account. In order to reduce the dimensionality of the original feature set, four feature selection techniques, namely, Consistency Based Analysis (CBA), Correlation Feature Set (CFS), Chi-Squared Attribute (CSA) and Principal Component Analysis (PCA) have been analyzed in Chapter 5. These feature selection techniques have been utilized to optimize the boundary extent features extracted from the word image. For classification purpose, two classifiers, namely, Decision Tree and Random Forest have been employed. In Chapter 6, Bootstrap Aggregating (Bagging) methodology has been considered to recognize offline handwritten Gurumukhi words based on four feature extraction techniques, namely, zoning features, centroid features, diagonal features and peak extent features (horizontally and vertically) along with their various combinations. Three classifiers, namely, k-NN, Decision Tree and Random Forest have been utilized to classify the words based on the extracted features. Chapter 7 introduces the concept of the eXtreme Gradient Boosting (XGBoost) technique in order to enhance the performance of offline handwritten Gurumukhi word recognition system based on four features, namely, zoning features, diagonal features, intersection & open-end points features, and peak extent features (horizontally and vertically). Finally, in Chapter 8, the conclusions have been drawn and the future directions based on this work have been mentioned.