Chapter 1

Introduction

1.1 Off-line Handwritten Word Recognition

Handwriting is one of the common modes of human communication. Handwritten document is basically an aggregated effect of intended textual information, writing style, writing pose and writing medium. Transcription of handwritten text into its digital form, is known as handwritten text recognition. In general, there are two types of handwritten text recognition: online recognition and off-line recognition. In the former case, the system recognizes written text while the writer writes, whereas in the latter case, the system is put to work after the completion of writing. Thus for online recognition, besides the written document, additional information such as pen tip movement, pen pressure, order of strokes writing, and stroke directions are also available; whereas for off-line handwritten word recognition the only available input data is the digital image of handwritten document. Therefore off-line handwritten text recognition is more difficult. This thesis presents a novel work on off-line handwritten word recognition. The advantage of handwritten text recognition is to involve a computer to process huge amount of handwritten data at high speed. Our research is focused on recognition of handwritten words written in globally used English script. In practice, three types of English handwritten scripts are available. They are (i) hand printed script (ii) cursive script and (iii) mixed script as shown in Fig.1.1. Visual appearance of same word in running handwriting (usually in cur-
Figure 1.1: Word written in (a) Hand printed script. (b) Cursive script. (c) Mixed script

Sive font) varies widely due to writer’s writing style, pose, mood and speed. An example is shown in Fig.1.2. Even this variation also may be present in same handwritten word written by the same writer (an example is shown in Fig.1.3). Therefore, automated handwritten word recognition is a very challenging task. Off-line handwritten word recognition has traditionally been tackled by following two main approaches: (i) Analytic approach and (ii) holistic approach. In the analytic approach a text-word is broken into subunits (e.g., characters, graphemes, allographs) and then each of these is identified. Finally, word recognition is performed by sequential analysis of the resulting subunit recognition scores. One of the main problems in this approach is that the decomposition of word into appropriate subunits is really a challenging job for cursive writing. On the other hand, the holistic approach avoids this decomposition step, and treats the text word as a single indivisible entity. In this case the features are extracted from the word as a whole. However, this approach is generally restricted to vocabulary of small size, i.e., where number of classes are small. In this thesis we have adopted the holistic approach for off-line handwritten word recognition.

1.2 Overview of handwritten word recognition system

Commonly handwritten word recognition system consists of the following steps: handwritten word extraction from a document page, preprocessing, feature ex-
1. INTRODUCTION

Figure 1.2: Visual appearance variation of same handwritten words written by five writers.

Figure 1.3: Visual appearance variation of same handwritten words written by same writer.

Figure 1.4: Basic steps of Handwritten text to digital text (taken from IAM database).
traction and recognition as shown in Fig.1.4. In practise, handwritten text may be written on an envelop/form/bank check/plain sheet. Thus, the first step of handwriting recognition is extraction of the region of interest. Sometimes prior knowledge about the document can help to locate the text zone. Subsequently the text line and finally, the text word are extracted from the text region. The second step is preprocessing, where word images are noise-cleaned, geometrically corrected and normalised. Commonly used preprocessing steps are skew correction, slope correction, slant correction and size normalisation. Then features are extracted from each word and the corresponding feature vector is formed. This feature vector represent handwritten words. The next step is recognition, where a suitable classifier is employed. The classifier recognises words and gives its decision. Sometimes for performance improvement, domain information is used in this step.

1.3 Literature review

Handwriting recognition is an interesting research topic since the 1950s [12, 113]. Handwritten document may be written in globally used English script [55] or in other regional scripts like Bengali, Arabic, Roman etc [10, 2, 16]. Generally handwritten word recognition is of two types, online recognition [9, 109, 6, 100] and off-line recognition [59, 98]. Researches on off-line handwritten word recognition focus on character recognition [5], digits recognition [84, 55, 33], text line recognition [8] and word recognition [11, 117, 13]. Handwritten word recognition system may be applied in restricted domains like address recognition and bank check processing or in less constrained type documents involving larger vocabularies [57, 11]. In this thesis we focus on off-line handwritten word recognition. Recognition of off-line handwritten text word has many applications such as reading addresses on postal pieces [105, 106, 30, 107], bank cheque processing [54, 47, 81, 123, 45, 87, 34, 108, 103, 88] and extracting data from various types of forms [73, 107]. Each steps of Off-line handwritten word recognition is explained in this dissertation.
1. INTRODUCTION

1.3.1 Database

Evaluation of a recognition system highly depends on the availability of appropriate databases. To establish a good, stable and acceptable recognition system, researchers need to test their algorithms on a number of databases. However, there are very few benchmark databases available for handwritten word recognition in English language. Mostly they are: CEDAR database[49] developed by Hull et al. in 1994, Cambridge database[101] developed by Senior et al. in 1998, CENPARMI database developed by Suen et al.[53], IAM database[77] developed by Marti et al. in 2002, C-Cube database developed by Camastra et al. in 2006[23]. CEDAR database contains 5000 city names, 5000 state names, 10000 ZIP Codes, and 50000 alphanumeric characters extracted from the digital images of handwritten addresses collected from the United States Postal Service (USPS). CENPARMI database is the English legal amount word database of CENPARMI, at Concordia University. The database has been created from 2500 handwritten cheques written by 800 different writers. Note that all writers had not contributed all words of lexicon. This database is partitioned into training set containing 2514 words and test set containing 5320 words. Note that there are different number of words in different word classes. On the other hand, IAM database contains handwritten English sentences and words. The Cambridge database is a collection of 4053 single writer words having a training set of 2675 words and a test set of 1378 words [68]. Cursive Character Challenge data set (C-Cube database) is an English character database, which can be downloaded from www.ccc.idiap.ch. This database contains cursive characters that are manually extracted from the CEDAR and United States Post Service (USPS) databases. A total 57293 character images including both uppercase and lowercase letters of English alphabet are split into: training set (38160 characters) and test set (19133 characters).

1.3.2 Preprocessing

In general, handwritten words are written in page or paragraph of text. To extract the words a document is segmented into text-line and then to handwritten words. Each handwritten word is finally normalised for feature extraction and
1. INTRODUCTION

recognition.

1.3.2.1 Text Line segmentation

In a database some writings are dark and some are faint due to variation of writing medium or digitization. To normalise the contrast variation binarization is a common approach. Binarization is the mapping of pixel values from $[0, 255]$ to $[0, 1]$ or [black, white]. Generally Otsu adaptive thresholding method [86] is used in literature [124, 11] for binarization. Some other researchers have used local binarization method [43, 121]. This method works well for old or degraded documents.

The next step is text line segmentation. Here we review few existing approaches to handwritten document line segmentation. They may be grouped as (i) projection based approach [75] (ii) smearing approach [65] and (iii) grouping approach [66]. Projection based approach and smearing approach are two popular line segmentation methods which are applicable to rectangular layout. A rectangular layout means that each element can be efficiently enclosed by a rectangular box. Such techniques are therefore suitable for cases where text lines are regular and parallel to a given direction. In projection based approach, the horizontal projection profile is obtained and then from this the vertical gap between two lines is determined and a line-separator is set at the center of this vertical gap. This approach works successfully if the gaps between two lines are significant and the lines are straight enough to resemble machine printed documents. For skewed line segmentation, this method is modified by Zabour et al. [4], Tripathy and Pal [114], Arivazhagan et al. [3] where the image of given text is divided into vertical strips. In smearing approach, consecutive black pixels along horizontal direction are smeared. This method requires an appropriate threshold for inter word gap. In grouping approach, text merging is done by aggregating units in bottom-up approach. Unit may mean pixel or connecting component or block. This technique cannot be used for degraded or poorly structured documents. Some researchers [42, 71, 67, 92] used Hough transform for text line segmentation. This approach is able to detect text lines in handwritten document which may contain lines oriented in different directions, erasures and annotation between main lines.
1.3.2.2 Slope and Slant correction

In an ideal example of handwriting, a word is supposed to be written horizontally with ascenders and descenders aligned along the vertical direction. In real data such conditions are rarely met. Slope (defined as, the angle between the horizontal direction and the direction of the implicit line (base line) on which the word is aligned) and slant (defined as, the angle between vertical direction and the direction of strokes supposed to be vertical) are often different than the desired value, zero. This difference ideally should be eliminated by slope and slant correction respectively.

Most of the desloping techniques presented in the literature are inspired by the method that tries to give an estimate of the core region (the region enclosing the main body of the characters) and then the base line on which the word is aligned is determined. Then the slope of the line is found and the image of word is rotated to get the desloped image. The core region is estimated by finding the lines with the highest horizontal density (number of foreground pixels per line). In general, the core region is denser than other regions (ascenders and descenders) of a word. Some researchers have estimated the core region by analyzing the distribution of density [80, 120, 63]. Lecolinet et al. have analyzed the entropy of the distribution (supposed to be lower when word is desloped), while Cai and Liu have worked on image rotation for each angle in an interval and the rotated image giving the highest peak of the first derivative of the horizontal density histogram was assumed as desloped [63, 21]. Some authors use piece-wise painting algorithm (PPA) for skew correction in [1]. Another popular method utilizes the eigen vectors of the covariance matrix of the row and column coordinates of the foreground pixels to find the axis of symmetry [25].

Most of the methods for slant correction are based on the selection of near vertical strokes, the slope of which is assumed as local slant estimate [21, 101]. The global slant value is obtained by averaging over all the local estimates. A different approach is followed by Vinciarelli and Luttin [120] where ‘deslantedness’ is measured over all the shear transforms of the word image corresponding to the angle in a reasonable interval. Some authors apply shear on a text-word by various angles and estimate the entropy [43] or variance [89] of the vertical projection of
1. INTRODUCTION

the text-word, and finally, select the angle for which the entropy is minimum or
the variance is maximum as the slant angle of the word.

1.3.2.3 Size normalization

Another important source of diversity in handwriting is writing size and thickness
of stroke. The size of ascenders and descenders also can vary from time to time.
Therefore before feature extraction, size of all words are resized to a common size
[97]. Some authors have first determined the ascenders and descenders present in
words and then scaled them into fixed height [89, 90]. Variation of pen and also
pen-pressure causes variation of stroke thickness. To normalise the thickness of
strokes, skeleton of black pixels is estimated which makes the stroke of one pixel
width [101]. Then using morphological dilation the stroke is made of a constant
width [60].

1.3.3 Feature extraction

Performance of a handwritten word recognition system depends mainly on proper
choice of feature set. Features that are used for word recognition can be grouped
into two levels: (a) low level and (b) high level.

Low level features are extracted from stroke fragments such as straight strokes,
curve strokes and bars, along with some other structural features like edges, end-
points and concavities. For example, stroke direction distribution is used as fea-
ture in [102, 48] for holistic recognition of hand-printed characters and machine-
printed words respectively. Bunke et al. [17, 18] have studied the percentage of
stroke-edges, present in ascender, descender and core region of a word and used
these as features. Percentage of foreground pixel in different region of a word is
taken as feature in [15, 82]. Vinciarelli [118] has applied sliding window method
for feature extraction, and number of foreground pixel present in each translated
window is taken as feature. Besides black pixel count [8], some other pixel based
features are pixels density [112, 11, 79], number of foreground/background tran-
sitions between adjacent cells [79], number of black and white pixels transition
[8] and position of center of gravity [79]. Another type of low level features is
1. INTRODUCTION

High level features used to recognize handwritten words are usually perceptual features [97, 72, 46]. Generally, in Holistic approach, handwritten words are recognised from their global shape. Thus most of the methods available in literature for cursive word recognition through holistic approach use high-level features. Guillevic and Suen [46] have used 11 components, namely the relative position of ascenders, descendents, loops and strokes as well as their numbers along with the word length, to form the feature vector. Cote et al. [31] have used three types of features to recognise cursive words. These are (i) primary features: ascenders, descenders, ascender-descenders and loop within the body of the word, (ii) secondary features: loops of 'b' and 'd' and the bar of 't', and (iii) face-up and face-down valley features. Kim et al. [53] have extracted shape feature, direction code distribution feature, curvature feature, moment feature and crossing feature to recognize legal amount words in bank cheque. Line orientation features and loop features have been used by Ruiz-Pinales et al. [97].

Besides traditional hand-crafted feature, some researchers are trying to learn features from training data using convolution neural network (CNN) [64, 126]. Researchers have started to use deeper architecture of CNN called deep CNN (DCNN) in various computer vision problems and are getting improved results compared to the state-of-the-art hand-crafted features [61, 83, 28, 14]. Some authors have used DCNN for feature extraction and the extracted features are recognised using different classifiers such as Support vector machine [83].

1.3.4 Word descriptor

Generally there are two approaches of word modeling: Modelling of whole word and modelling of character or sub-word. In modelling of whole word [31, 74], feature is extracted directly from word as a whole without segmentation. The aim of this model is to extract overall shape information like word contour shape or presence of ascenders, descendents, loops etc. In this model segmentation can be avoided but poses limitation on lexicon size. In deep convolutional neural
networks [64, 126, 83, 28, 14], 2D image is convolved with several filters and resized using sub-sampling operation in several layers. In modelling of character or sub-word, first the image is segmented into characters or sub-words, which is a truly challenging job for off-line handwritten words. Segmentation of word to sub-words may be done using control points [37], regularity and singularity principles [56], local minima of upper contour [81]. Tay et al. [110] had segmented words into horizontal slices depending on the height of core region. This model suffers from over-segmentation of image. Without relying on heuristics segmentation approach, the segmentation free model becomes most popular nowadays. One commonly used strategy of this model is sliding window method [118, 51].

1.3.5 Classification

A large variety of classifiers have been employed for handwritten word recognition. Encouraged by the success in automated speech recognition, Hidden Markov Model (HMM) has been widely applied to off-line handwriting recognition [39, 91, 78, 117, 50, 119, 44]. Within the HMM framework, there are several ways to introduce context while modeling individual words. Some authors have dealt with contextual modeling in handwritten word recognition [41, 38]. In many works, artificial neural networks have been applied to classify characters as part of continuous handwritten word [20, 19, 58, 111, 76] and cursive characters [24, 22, 32, 115, 94]. Some methods use K-nearest neighbour (KNN) classifier [62, 108]. Popular support vector machine (SVM) classifiers are also used by various systems for handwritten cursive character recognition [22, 94]. In fact, Rehman and Verma [94] have presented a rigorous experimental study using various feature sets and three different classifiers: NN, KNN and SVM. They have discussed advantages and disadvantages of different feature sets as well as classifiers. To improve performance, researchers have focused on hybrid classifiers where results of multiple classifiers are combined [47, 53, 54, 40, 32, 97]. Vamvakas et al. [115] follow hierarchical recognition scheme to combine decision of all classifiers.
1.4 Motivation and contributions of this thesis

Interpreting handwritten document by computer is a challenging task and is an active research topic. Developing an innovative algorithm to match human recognition ability is a challenging problem. Though numerous scientists have been working on this subject for many years, there is still a long way to go. Intensive research on the recognition of isolated digits in the past decade has led to very high recognition rates in the range of 95-99%. But the recognition of cursive words is still far below satisfaction. For small lexicon words, the recognition rate is in the range of 85-93%. Far from being solved, off-line handwritten word recognition still interests of researchers. Thus it is always interesting to explore a new technique and attempt to improve off-line handwritten word recognition. This research is focused on developing novel algorithms to recognise off-line handwritten words. The main contributions of this thesis are as follows.

- The thesis presents a new ensemble classifier based off-line handwritten word recognition system. Here we have followed a holistic approach for feature extraction. Thus our system is restricted to small lexicon. The system is tested on three handwritten word databases. The performance of proposed system is promising and comparable to state of art handwriting recognition system.

- A new handwritten word database, named ISIHWD is proposed. This database contains 31124 handwritten words written by 105 different writers. Fixed partition for training and test set is given to provide common platform to all researchers. ISIHWD contains 33 different words which are legal words mainly used in Indian bank cheques.

- A novel model based text line segmentation method is proposed. The basic concept behind this segmentation method is that two successive handwritten text-lines are always non-intersecting. This method works well for horizontally written text documents.

- Two novel slope correction techniques of handwritten words are proposed. In the first approach, the base line of word is estimated from the centroids
1. INTRODUCTION

of narrow vertical strips of core region of word using leverage, whereas in the second approach the distribution of black pixels of text word image is exploited.

- Three novel slant correction techniques of handwritten words are proposed where Gabor filter, Fast Fourier transform and Hough transform are used respectively to detect the slant angle of word.

- A new Arnold transform based directional feature is proposed. Here direction of strokes present in word are estimated using Arnold transform followed by Hough transform. The performance of proposed feature on handwritten words is moderately good.

- A new curvelet based feature is proposed. Here a word is decomposed into small arcs. Each arc is labelled using Fibonacci number scheme to ensure uniqueness. The histogram of curvelet index is taken as main feature. The word recognition accuracy of this feature is good for handwritten word recognition.

- A new architecture of Deep convolution neural network (DCNN) is proposed for handwritten word recognition. Here we have used the DCNN to extract features of word. The DCNN based features are very good representation of handwritten word.

1.5 Organization of this thesis

This thesis is divided into seven Chapters. The current chapter contains a brief literature review along with aims and content of the thesis.

Chapter 2 gives a detail description of off-line handwritten word databases used to evaluating the performance of the proposed algorithms. Section 2.1 describes two benchmark handwritten word databases: i) CENPARMI database and ii) IAM database. Section 2.2 gives a detailed description of ISIHWD database developed by us. In Section 2.3, basic preprocessing steps are explained where novel text line segmentation method as well as slope and slant correction techniques are proposed.
Chapter 3 gives the detailed description of handwritten word recognition technique using Arnold Transform based features. A brief description of Arnold transform is given in Section 3.1. In Section 3.2 Arnold transform based features and naive directional features are explained. The experimental results are described in Section 3.3. Finally, the proposed algorithm is compared with existing systems in Section 3.4.

In Chapter 4, a new handwritten word recognition technique is proposed using oriented curvelet feature. The proposed method is described in Section 4.1 where pixel based curvelet index computation using Fibonacci number scheme is explained. Details of the experimental results and comparison with existing systems are given in Section 4.2 and Section 4.3 respectively.

Chapter 5 describes handwritten word recognition technique using DCNN. Proposed architecture of DCNN is explained in Section 5.2 after introducing a brief description of DCNN in Section 5.1. Detail experimental results and comparison with existing systems are given in Section 5.4 and Section 5.5 respectively.

In Chapter 6, an Ensemble Classifier based Off-line Handwritten Word Recognition System is presented. This hybrid handwritten word recognition system is explained in Section 6.1. The combining scheme is described in Section 6.2. Finally, experimental results and comparative study are discussed in Section 6.3 and 6.4 respectively.

The final chapter 7 draws the conclusion of the proposed work and provides a brief summary of future scope of research in this area.